Toward Metrics for Differentiating Out-of-Distribution Sets

Mahdieh Abbasi[1], **Changjian Shui**[2], **Arezoo Rajabi**[3], **Christian Gagné**[4] and **Rakesh B. Bobba**[5]

Abstract. Vanilla CNNs, as uncalibrated classifiers, suffer from classifying out-of-distribution (OOD) samples nearly as confidently as in-distribution samples. To tackle this challenge, some recent works have demonstrated the gains of leveraging available OOD sets for training end-to-end calibrated CNNs. However, a critical question remains unanswered in these works: how to differentiate OOD sets for selecting the most effective one(s) that induce training such CNNs with high detection rates on unseen OOD sets? To address this pivotal question, we provide a criterion based on generalization errors of Augmented-CNN, a vanilla CNN with an added extra class employed for rejection, on in-distribution and unseen OOD sets. However, selecting the most effective OOD set by directly optimizing this criterion incurs a huge computational cost. Instead, we propose three novel computationally-efficient metrics for differentiating between OOD sets according to their "protection" level of in-distribution sub-manifolds. We empirically verify that the most protective OOD sets – selected according to our metrics – lead to A-CNNs with significantly lower generalization errors than the A-CNNs trained on the least protective ones. We also empirically show the effectiveness of a protective OOD set for training well-generalized confidence-calibrated vanilla CNNs. These results confirm that 1) all OOD sets are not equally effective for training well-performing end-to-end models (i.e., A-CNNs and calibrated CNNs) for OOD detection tasks and 2) the protection level of OOD sets is a viable factor for recognizing the most effective one. Finally, across the image classification tasks, we exhibit A-CNN trained on the most protective OOD set can also detect black-box FGS adversarial examples as their distance (measured by our metrics) is becoming larger from the protected sub-manifolds.

1 Introduction

In supervised learning, it is generally assumed that a training set and a held-out test set are drawn independently from the same data distribution, called *in-distribution set*. While this assumption can be true for controlled laboratory environments, it rarely holds for many real-world applications, where the samples can be drawn from both in-distribution and from other distributions, called *out-of-distribution* (OOD) data, which contains samples that are semantically and statistically different from those in-distribution. In the presence of OOD samples, it is important to have a model able to distinguish them in order to make reliable decisions. However, it has been shown that state-of-the-art (vanilla) deep neural networks (e.g., CNN) are uncalibrated such that they are making predictions for OOD samples with a confidence that is as high as those of in-distribution samples, making them indistinguishable from each other [12, 13]. For safety-critical real-world applications such as self-driving cars, using vanilla CNNs that tend to confidently make wrong decisions for such unknown OOD samples can lead to serious safety and security consequences.

To tackle this challenge, post-processing approaches [7, 14, 20, 21] attempt to transform the confidence of predictions made by pre-trained vanilla CNNs in order to create a gap between the confidence for OOD samples and that of in-distribution ones. Despite their simplicity and efficiency, their performances depend on several additional hyper-parameters such as temperature, magnitude of additive noise, or the parameters of an auxiliary regression function, which should be carefully tuned for each OOD set.

More recently, some researchers [24, 4, 15, 19, 23] have proposed *end-to-end* calibrated CNNs-based models for OOD detection. For instance, calibrated vanilla CNNs [15, 13, 20] are trained to make uncertain predictions for OOD samples while still confidently and correctly classifying in-distribution ones. To train these models, some authors [24, 4, 15, 23] have leveraged a naturalistic OOD set[6], which seemingly is selected *manually* from among many available ones, without providing a systematic justification for their selection. Thus, the following question remains unaddressed in these works: *how to differentiate among OOD sets w.r.t. a given in-distribution task with the goal of selecting the most proper one, which in turn induces a well-generalized calibrated model with high detection rate of* unseen *OOD sets?*.

Besides the confidence-calibrated vanilla CNN [24, 15, 19] as end-to-end model for OOD detection task, the classical idea of adding an explicit rejection class [1, 2, 6, 11] is also an interesting end-to-end approach. Indeed, such augmented classifiers can directly reject OOD samples by classifying them to the extra class, while correctly classifying in-distribution samples. In addition to the calibrated vanilla CNN, we exploit A-CNN as an end-to-end model for OOD detection task.

However, without having a principle for selecting the right OOD sets among those available, training *well-generalized A-CNN and calibrated vanilla CNN* is challenging. Since using

---

[1] Université Laval, Canada, `mahdieh.abbasi.1@ulaval.ca`
[2] Université Laval, Canada, `changjian.shui.1@ulaval.ca`
[3] Oregon State University, USA, `rajabia@oregonstate.edu`
[4] Canada CIFAR AI Chair / Université Laval, Canada, `christian.gagne@gel.ulaval.ca`
[5] Oregon State University, USA , `rakesh.bobba@oregonstate.edu`

---

[6] We simply drop naturalistic and instead call it OOD set throughout the paper.

a randomly selected OOD set does not necessarily lead to a model with a high detection rate of unseen OOD sets (i.e., generalization ability) as we later show in our experiments. It has also been shown in [19], where using SVHN as OOD set for CIFAR-10 is leading to an A-CNN with inferior generalization properties. Moreover, simply using a union of an enormous number of OOD sets not only creates an extremely unbalanced dataset, but also makes training of these models computationally infeasible.

Although Hendrycks et al. [15] have conjectured diversity for characterizing a proper OOD set, in this paper, our main focus is to answer concretely the aforementioned question, i.e., how to differentiate between OOD sets in order to select a proper one. At first, we provide a formal criterion in the form of generalization errors of A-CNN for differentiating OOD sets and selecting the most effective one. Using this, an OOD set is recognized as a proper (effective) if it leads to training of A-CNN with low generalization errors for both in-distribution and unseen OOD sets. However, directly optimizing this selection criteria is computationally very expensive. To overcome this, we propose some metrics that can be efficiently computed using a pre-trained vanilla CNN. We drive our metrics according to the following intuition: a proper (effective) OOD set should cover sub-manifolds of an in-distribution task, which can be achieved by the penultimate layer of a vanilla CNN trained on it. Thus, we design our metrics to measure the degree of protectiveness of sub-manifolds by OOD sets for selecting the most protective one. Indeed, protecting in-distribution sub-manifolds by OOD samples allows for rejecting *automatically* the unseen OOD sets which are located relatively far away from the protected in-distribution sub-manifolds, as shown in Figure 1. Therefore, the protection level of OOD sets can be a viable factor for differentiating and selecting of OOD sets with the aim of obtaining a well-generalized A-CNN and calibrated vanilla CNN.

Our contributions in this paper can be outlined as follows:

- We provide a formal definition (with the use of A-CNN) for precisely differentiating OOD sets according to their induced generalization errors on unseen OOD sets and in-distribution set.
- We are first to propose novel quantitative metrics for differentiating OOD sets with the aim of selecting the most protective OOD set w.r.t. a given in-distribution set. These metrics, namely **Softmax-based Entropy (SE)**, **Coverage Ratio (CR)** and **Coverage Distance (CD)**, can be *efficiently* computed using a vanilla CNN trained on the given in-distribution task.
- In an extensive series of experiments on image and audio classification tasks, we empirically show that A-CNNs and calibrated vanilla CNNs trained on the most protective OOD set have higher detection rates (lower generalization error) on unseen OOD sets in comparison with those trained on the least protective OOD set.
- We exhibit that A-CNN trained on the most protective OOD set (i.e., A-CNN*) can also detect black-box FGS adversarial examples generated by a relatively large magnitude of noise, while vanilla CNN and the A-CNN trained on the least protective OOD set are still incorrectly classifying them. We show this occurs as the distance (measured by CD) of FGS adversaries is increased from the protected sub-manifolds.

2 Characterizing a Proper OOD Set

Let us assume a hypothesis class $\mathcal{H}'$ (e.g. A-CNN) for a $K+1$ classification problem with K classes associated for a given in-distribution task and the extra class (i.e., $(K+1)$-th class) reserved for identifying OOD samples. We also denote $\mathcal{S}_I = \{(\mathbf{x}_I^i, \mathbf{y}_I^i)\}_{i=1}^N$ as an in-distribution training set consisting of N i.i.d. labeled samples drawn from data distribution $\mathcal{D}_I$, with true labels $\{\mathbf{y}_I^i\}_{i=1}^N \in \{1, \ldots, K\}$. As the OOD training set, take $\mathcal{S}_O = \{(\mathbf{x}_O^j)\}_{j=1}^M$ involving M i.i.d. samples drawn from a data distribution $\mathcal{D}_O$, which we label as $(K+1)$-th class.

The loss of a hypothesis $h' \in \mathcal{H}'$ for a given in-distribution sample can be defined as $\ell(h'(\mathbf{x}_I^i), \mathbf{y}_I^i) = \mathbb{I}(h'(\mathbf{x}_I^i) \neq \mathbf{y}_I^i)$ and its loss for an OOD sample is $\ell(h'(\mathbf{x}_O^j), K+1) = \mathbb{I}(h'(\mathbf{x}_O^j) \neq K+1)$[7]. The *true* loss of an augmented classifier $h' \in \mathcal{H}'$ can be evaluated on the underlying data distributions $\mathcal{D}_I$ and $\mathcal{D}_O$ as:

$$L_{\mathcal{D}_I}(h') = \mathbb{E}_{(\mathbf{x}_I, \mathbf{y}_I) \sim \mathcal{D}_I} \ell(h'(\mathbf{x}_I), \mathbf{y}_I), \qquad (1)$$

$$L_{\mathcal{D}_O}(h') = \mathbb{E}_{\mathbf{x}_O \sim \mathcal{D}_O} \ell(h'(\mathbf{x}_O), K+1). \qquad (2)$$

The corresponding *empirical* loss is computed on training set $\mathcal{S}_I$ and $\mathcal{S}_O$:

$$L_{\mathcal{S}_I}(h') = \frac{1}{N} \sum_{i=1}^N \ell(h'(\mathbf{x}_I^i), \mathbf{y}_I^i), \qquad (3)$$

$$L_{\mathcal{S}_O}(h') = \frac{1}{M} \sum_{j=1}^M \ell(h'(\mathbf{x}_O^j), K+1). \qquad (4)$$

Before presenting our definition, we remark that there is a set of B "out" data distributions $\mathcal{D}_O^b, b = \{1, \ldots, B\}$ with their respective OOD training set $\mathcal{S}_O^b \sim \mathcal{D}_O^b$. Theoretically speaking, B can be infinitely large. Moreover, we assume generalization error of vanilla classifier (denoted by $h \in \mathcal{H}$), for the original K classification task, trained on $\mathcal{S}_I$ is less than a small ϵ value: $|L_{\mathcal{S}_I}(h) - L_{\mathcal{D}_I}(h)| \leq \epsilon$.

Definition 1 : *For a given OOD training set $\mathcal{S}_O^b \sim \mathcal{D}_O^b$ and in-distribution training set $\mathcal{S}_I$ w.r.t. hypothesis class $\mathcal{H}'$, $\mathcal{D}_I$ and B "out" data distributions, we define two kinds of gaps for the augmented classifier $h_b' \in \mathcal{H}'$ trained on $\mathcal{S}_I \cup \mathcal{S}_O^b$, i.e. $\min_{h_b'} L_{\mathcal{S}_I} + L_{\mathcal{S}_O^b}$:*

$$\boldsymbol{\mathcal{L}}_{\mathcal{S}_I} = |L_{\mathcal{S}_I}(h_b') - L_{\mathcal{D}_I}(h_b')|, \qquad (5)$$

$$\boldsymbol{\mathcal{L}}_{\mathcal{S}_O^b} = \sup_{\mathcal{D}_O \in \{\mathcal{D}_O^1, \ldots \mathcal{D}_O^B\}} |L_{\mathcal{S}_O^b}(h_b') - L_{\mathcal{D}_O}(h_b')|. \qquad (6)$$

The first term $\boldsymbol{\mathcal{L}}_{\mathcal{S}_I}$ represents the gap between empirical loss of classifier $h_b' \in \mathcal{H}'$ on in-distribution training set $\mathcal{S}_I$ and its true loss on $\mathcal{D}_I$ while the second term $\boldsymbol{\mathcal{L}}_{\mathcal{S}_O^b}$ concerns the largest (worst) gap between empirical loss of h_b' on OOD training set $\mathcal{S}_O^b$ and its true loss on "out" data distributions. By restricting B to a manageable (finite) large number, we re-define $\boldsymbol{\mathcal{L}}_{\mathcal{S}_O^b}$ by upper-bounding Eq. 6, i.e. sum of gaps on B finite "out" data distributions:

$$\boldsymbol{\mathcal{L}}_{\mathcal{S}_O^b} = \sum_{\mathcal{D}_O \in \{\mathcal{D}_O^1, \ldots \mathcal{D}_O^B\}} |L_{\mathcal{S}_O^b}(h_b') - L_{\mathcal{D}_O}(h_b')|. \qquad (7)$$

---

[7] Indicator function $\mathbb{I}(p)$ returns 1 if condition p is true, and 0 otherwise.
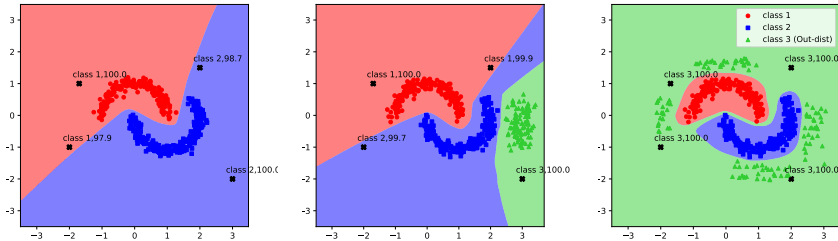
Figure 1: Illustration of properties of a partially-protective OOD set (middle) and a protective one (right) and their effect on training A-MLP for a two-moon classification dataset, compared to a (left) vanilla MLP trained on the same dataset. The black-cross samples are some test OOD samples and their predicted class and confidence scores by each (A)-MLP are also indicated. All MLPs are composed of three layers with ReLU activation function.

As true data distributions are unknown, the aforementioned equations can be empirically computed using validation sets. Then, a proper OOD set is the OOD set that training a A-CNN on it should produce the lowest accumulation of generalization errors of both in-distribution task and (un)seen OOD sets:

$$\mathcal{S}_O^{b^*} = \underset{\mathcal{S}_O^b \in \{\mathcal{S}_O^1, \cdots \mathcal{S}_O^B\}}{\operatorname{argmin}} \mathcal{L}_{\mathcal{S}_I} + \lambda \mathcal{L}_{\mathcal{S}_O^b}, \qquad (8)$$

where $\lambda > 0$ is a balancing hyper-parameter. Directly using Eq. 8 to find a proper OOD set is computationally inefficient as it involves training B individual augmented classifiers, i.e. train each h_b' on a pair of $\mathcal{S}_I \cup \mathcal{S}_O^b, b \in \{1, \ldots, B\}$. Particularly for the case of CNNs, this incurs a huge computational overhead. To overcome this computational burden, we conjecture that a protective OOD set can also provide a well-generalized A-CNN on both in- and unseen OOD sets (an intuitive illustration is given in Sec. 2.1). Thus, instead of directly optimizing Eq 8, we develop some cost-effective metrics to assess the protectiveness level of OOD sets for identifying the most protective one.

2.1 Protective OOD set: An Illustrative Example

To give a high-level intuitive explanation of our proposed metrics for recognizing a protective OOD set, we use an example based on the two-moon dataset (as in-distribution task), where each moon is considered as a sub-manifold. Fig. 1(a) exhibits the challenge of OOD samples for a vanilla MLP, which is trained on only in-distribution samples. As can be seen, this vanilla MLP *confidently* classifies OOD samples (indicated with black-crosses) as either "class 1" or "class 2" albeit they clearly belong to *none* of the in-distribution manifolds. In Fig. 1(b) we demonstrate a *partially-protective* OOD set whose samples are almost collapsed and only partially cover one of the sub-manifolds (i.e., the manifold with blue squares). An augmented MLP (A-MLP) trained on the two-moon dataset along with this OOD set leads to a classifier with a limited OOD detection performance (lower generalization ability of detecting unseen OOD samples). More precisely, OOD samples, e.g., the *unseen* black-cross samples, which are laying around *uncovered* parts of the manifolds, are still confidently misclassified by the underlying A-MLP. Whereas, in Fig 1(c) a proper *protective* OOD set, whose samples better cover the in-distribution's sub-manifolds (two-moon), is shown. As can be seen, training an A-MLP on such a protective OOD set (along with in-distribution samples) leads to classifying *unseen*

black cross OOD samples as class 3 (i.e., the extra class) as well as classifying automatically the regions out of the manifolds as class 3. This results in an A-MLP with high detection performance on unseen OOD sets (i.e., making the gap in Eq. 6 small). Therefore, the design of our metrics is driven according to this intuition that a proper OOD set should be more protective of (i.e., closely covers) all in-distribution sub-manifolds in the feature space. A similar intuition has been previously exploited by some researchers, e.g. [19, 33], with the aim of generating synthetic OOD samples.

2.2 Proposed Metrics

As previously done [8, 16], we consider the penultimate layer of a vanilla CNN as a function that transfers samples from high-dimensional input space into a low-dimensional feature space, placing them on data (sub-)manifold(s) [3, 5]. Furthermore, we assume that for a standard multi-classification problem (with K classes), each class has its own sub-manifold in the feature space where its associated in-distribution samples lie. In the following, we propose our metrics to assess which of the available OOD sets has a better and closer coverage of the sub-manifolds.

2.2.1 Softmax-based Entropy

Our first metric aims at determining whether the samples of a given OOD set are distributed *evenly* to all sub-manifolds (of a given in-distribution task) such that they have the equal chance of being covered by these OOD samples. For example, an OOD set, whose samples are misclassified by a given vanilla CNN into *only* a few of in-distribution classes (manifolds) instead of all of them, is deemed as a non-protective OOD set. This is because the sub-manifolds with no or only a few OOD samples being misclassified to them, are still uncovered (cf. Fig. 1(b)). Thus training A-CNN on such non-protective (or partially-protective) OOD set may lead to limited detection performance of unseen OOD sets. In contrast, the samples of a protective set are expected to be misclassified evenly to all the sub-manifolds, giving all of them an equal chance of being covered.

To quantitatively measure this incidence for a given OOD set w.r.t. an in-distribution set and a vanilla CNN trained on it, we introduce *Softmax-based Entropy* (SE). First we define $p(c = k | \mathcal{S}_O)$ as the conditional probability of k-th class given

$\mathcal{S}_O$ as follows:

$$p(c = k|\mathcal{S}_O) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}_k \left(\text{argmax}(h(\mathbf{x}_O^j)) = k \right), \quad (9)$$

where h is softmax output of the vanilla CNN trained on $\mathcal{S}_I$ and $\mathbb{I}_k(.)$ is an indicator function for k-th class. It returns 1 if a given OOD sample $\mathbf{x}_O^j \in \mathcal{S}_O$ is (mis)classified as class k by h, otherwise it returns 0. Finally, SE is defined for an OOD set $\mathcal{S}_O$ as follows:

$$H(\mathcal{S}_O) = -\sum_{k=1}^{K} p(c = k|\mathcal{S}_O) \log p(c = k|\mathcal{S}_O). \quad (10)$$

$H(S_O)$ shall reflect how uniformly the samples of $\mathcal{S}_O$ are distributed to in-distribution sub-manifolds (i.e., corresponding to each in-distribution class). Note that the maximum value of SE is $\log K$ (K is the number of classes) when all of the samples are uniformly distributed. Thus, the highest $H(\mathcal{S}_O)$ indicates that all the sub-manifolds have an equal number of OOD samples, whereas the smallest value of $H(\mathcal{S}_O)$ indicates some sub-manifolds (except a few of them) have a small number of (or no) OOD samples to cover them. Therefore, a protective OOD set should have a higher SE than that of non-protective ones.

2.2.2 Coverage Ratio

Although an OOD set with the high(est) SE confirms OOD samples are evenly distributed to all the sub-manifolds, using solely SE is not sufficient to assure the coverage of these sub-manifolds. Putting differently, an OOD set with the highest SE might still be collapsed and only partially cover some parts of the sub-manifolds.

Inspired by covering number notion [29], we introduce our second metric, named coverage ratio (CR), in order to measure coverage of the sub-manifolds. Recall the sub-manifolds are approximated using a training in-distribution set in the feature space that is achieved by the penultimate layer of h. We denote $\mathbf{z}_I^i$ and $\mathbf{z}_O^j$ as the representations of $\mathbf{x}_I^i \in \mathcal{S}_I$ and $\mathbf{x}_O^j \in \mathcal{S}_O$ in the feature space, respectively.

To formally describe Coverage Ratio (CR), we form a rectangular weighted adjacency matrix $W \in \mathbb{R}^{N \times M}$ for a given pair $(\mathcal{S}_I, \mathcal{S}_O)$ with N in-distribution and M OOD samples, respectively. $W_{i,j} = \|\mathbf{z}_I^i - \mathbf{z}_O^j\|_2$ is the distance (l_2-norm) between in-distribution sample $\mathbf{z}_I^i$ and OOD sample $\mathbf{z}_O^j$ in the feature space. The distance between a pair of $(\mathbf{z}_I^i, \mathbf{z}_O^j)$ is computed only if $\mathbf{z}_I^i$ is among k-nearest in-distribution neighbors of $\mathbf{z}_O^j$, otherwise $W_{i,j} = 0$:

$$W_{i,j} = \begin{cases} \|\mathbf{z}_I^i - \mathbf{z}_O^j\|_2 & \text{if } \mathbf{z}_I^i \in \text{k-NN}(\mathbf{z}_O^j, S_I) \\ 0 & \text{otherwise} \end{cases}. \quad (11)$$

In other words, for each sample $\mathbf{z}_O^j$, we find its k-nearest neighbors from the in-distribution set $\mathcal{S}_I$ in *the feature space*. Then, if the given $\mathbf{z}_I^i$ belongs to k-nearest in-distribution neighbors of $\mathbf{z}_O^j$, we set $W_{i,j}$ to their distance. From matrix W, we derive a binary adjacency matrix A as follows; $A_{ij} = \mathbb{I}(W_{ij} > 0)$. Now using matrix A, we define CR metric as follows:

$$R(\mathcal{S}_I, \mathcal{S}_O) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left(\sum_{j=1}^{M}(A_{i,j}) > 0 \right), \quad (12)$$

where $\mathbb{I}(\sum_{j=1}^{M}(A_{i,j}) > 0)$ assesses whether the i-th in-distribution sample $\mathbf{z}_I^i$ covered at least one time by the OOD samples $\mathcal{S}_O$ in the feature space. Basically, this metric measures how many in-distribution samples (percentage) are covered by at least one OOD samples from $\mathcal{S}_O$ in the feature space. Finally, **we estimate an OOD set w.r.t. a given in-distribution set is protective if it has both high SE and high CR.**

It is important to note that SE and CR are complementary. As mentioned earlier, high SE of an OOD set without considering its CR is not sufficient for estimating the protective level of an OOD set. Similarly, from high CR alone without having high SE, an OOD set cannot be considered as a protective one. This is because, an OOD set with high CR but low SE is not distributed enough among all sub-manifolds and might cover a large portion of only a few sub-manifolds.

2.2.3 Coverage Distance

Furthermore, to measure the distance between OOD set $\mathcal{S}_O$ and the in-distribution data sub-manifolds, the following distance metric, named Coverage Distance (CD), can be driven:

$$D(S_I, S_O) = \frac{\sum_{i,j} W_{ij}}{\sum_{i,j} A_{ij}} = \frac{1}{kM} \sum_{i,j} W_{ij}. \quad (13)$$

$D(\mathcal{S}_I, \mathcal{S}_O)$ shows average distance between OOD samples of $\mathcal{S}_O$ and their k nearest neighbors from in-distribution set. **Selection of Protective OOD set**: We remark that for final selection OOD set *SE and CR play more important roles than CD since they indicate the degree of spread and protectiveness of OOD sets for the sub-manifolds while CD reveals the average distance of OOD set to the sub-manifolds.* Since our primary concern is about coverage of the sub-manifolds, we first assess SE and CR. In other words, the most protective OOD set should have the highest SE (preferably near to $\log K$) and the highest CR compared to those of the remaining OOD sets. If one encounters some OOD sets that have (relatively) equal highest SE and CR, then their CDs can be considered for final selection –the OOD set with smaller CD can be selected.

3 Experimentation

We conduct a series of experiments on several classification tasks including two image benchmarks, namely CIFAR-10 and SVHN, and one audio benchmark, namely Urban-Sound [28]. In our experiments, we utilize VGG-16 and a CNN described in [27] for image and audio benchmarks, respectively.

Like in [21], for each of these in-distribution task, various naturalistic OOD sets are considered; for image classification tasks, we consider LSUN, ISUN, CIFAR-100 and TinyImageNet as OOD sets and Gaussian noise as a synthetic OOD set. For audio classification task with 10 classes, i.e., Urban-Sound, OOD sets considered are TuT [25], Google Command [32] and ECS (Environmental Sound Classification) [26], as well as white-noise sound as a synthetic OOD set. Note the classes of an OOD set, which are semantically or exactly overlapping with those of the given in-distribution set, are discarded.

In our experiments, we consider two types of end-to-end approaches, i.e., an A-CNN and a confidence-calibrated vanilla CNN, for detecting OOD set. The latter type (i.e., calibrated vanilla CNN), a CNN is said to be calibrated after being trained
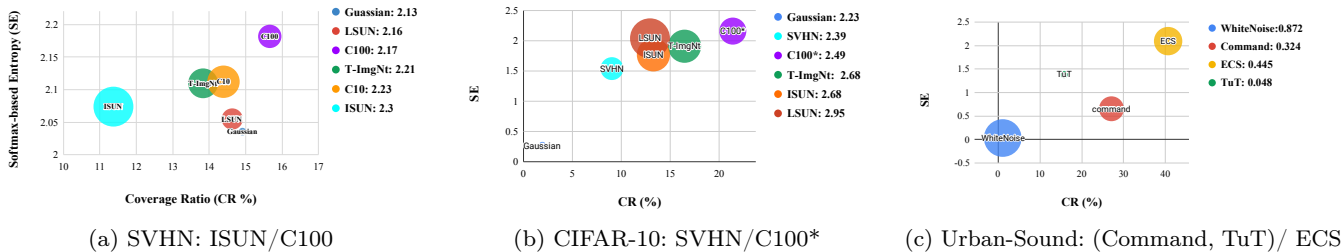
| (a) SVHN: ISUN/C100 | (b) CIFAR-10: SVHN/C100* | (c) Urban-Sound: (Command, TuT)/ ECS |

Figure 2: Differentiating OOD sets for SVHN, CIFAR-10, and Urban-Sound for the purpose of selecting the most protective one using our proposed metrics. Each sub-figure shows a bubble chart with SE and CR as y-axis and x-axis, respectively. The size of bubbles is determined by CD, also shown in the legend of the sub-figures. The least/most protective OOD sets are indicated in caption of the sub-figures.

to predict OOD training samples with great uncertainty (i.e. uniform prediction) while confidently classifying correctly in-distribution training samples. To achieve this, instead of cross entropy loss, we apply the modified loss function used in [15, 19]. As the calibrated CNNs are threshold-based models for performing OOD detection, likewise [15, 21, 20] to assess its performance, we report AUROC, FPR at 95% TPR on test for in- and out-of-distribution sets.

As A-CNN has an explicit rejection option (i.e. threshold-free approach), we consider three criteria to assess its performance on the in-distribution test set: 1) Accuracy rate (Acc.) ↑, that is the rate of samples classified correctly as their true associated label (↑ indicates the higher the better), 2) Rejection rate (Rej.) ↓, the rate of samples misclassified as dustbin (↓ indicates the lower the better), 3) Error rate (Err.) ↓, therate of samples that are neither correctly classified nor rejected (Err. rate = 1- (Acc. rate + Rej. rate)). A-CNN performance on OOD sets is evaluated by I) Rejection rate (Rej.) ↑: percentage (rate) of OOD samples classified as dustbin, and II) Error rate (Err.) ↓: rate of OOD samples not classified as dustbin (Err. = 1 - Rej.) Note since A-CNNs is a threshold-free OOD detector there are no AUROC and AUPR values. Plus, A-CNN's OOD rejection rate and its in-distribution rejection rate are the same concepts as TNR (True Negative Rate) and FNR (False Negative Rate), respectively.

3.1 Empirical Assessment of Metrics

First, to obtain in-distribution sub-manifolds, if in-distribution training set has more than $10,000$ samples, we randomly select $10,000$ samples from it, otherwise, we use the whole in-distribution training set. Secondly, we pass the samples through the penultimate layer of a pre-trained vanilla CNN to map them into the feature space. The same procedure is done for the samples of an OOD set to transfer them to the feature space. To fairly compare OOD sets according to the metrics, we also randomly select equal number of OOD samples ($10,000$ samples) from each OOD set. For OOD sets with various sizes, we take the minimum size for making equal size OOD sets by randomly selecting from them. To compute CR and CD metrics, we set our only hyper-parameter, i.e. the number of nearest neighbors, $k = 4$ for all our experiments. Note that among our metrics, CR and CD are dependent on k (the impact of k on our metrics are presented later).

3.1.1 Differentiating OOD Sets by the Metrics

Using the proposed metrics, we differentiate OOD sets for each in-distribution set to select the most protective OOD sets w.r.t. the given in-distribution. In Fig. 2, we demonstrate the difference between OOD sets according to their SE, CR, and CD, in order to identify the most and least protective OOD set. The most and least protective naturalistic OOD sets identified by our metrics (particularly by SE and CR) are indicated in caption of sub-figures in Fig. 2. For *SVHN* task, for example, *ISUN*, among naturalist OOD sets, and *Gaussian noise*, as synthetic OOD set, are identified as the least protective sets. Note that despite the high CR of Gaussian noise, its SE is far small, indicating it as a collapsed OOD set, which thus causes it to be identified as the least protective. The most protective OOD set for SVHN is CIFAR-100 (i.e., C100) with the highest SE and CR. For Urban-sound, ECS is the most protective, while Command and TuT datasets can be regarded as non-protective OOD sets due to their significantly low SE with respect to the upper bound of SE ($\log 10$).

To assess the sensitivity of our metrics to the choice of k, we show the CR of OOD sets for varying k's values in Fig 3. In our experiments, we observe that the relative ranking of OOD sets according to their CR and CD is consistent with various values of k. Thus, CR and CD are not sensitive to the choice of k.
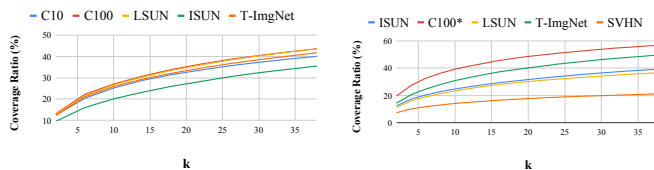

Figure 3: The effect of k (number of nearest neighbors) on CR of OOD sets for SVHN (left) and CIFAR-10 (right).

A-CNN: As Eq. 8 states, training an A-CNN on a proper OOD set should lead to a low average of [8] error rates on (un)seen OOD sets (or equivalently high average of OOD sample rejection rates). Therefore, for a given in-distribution task (e.g. SVHN), we train a separate A-CNN on each OOD set. Then, the error rates of these A-CNNs on all of the OOD sets are evaluated. Note a small error rate on a OOD set is equivalent to high detection rate.

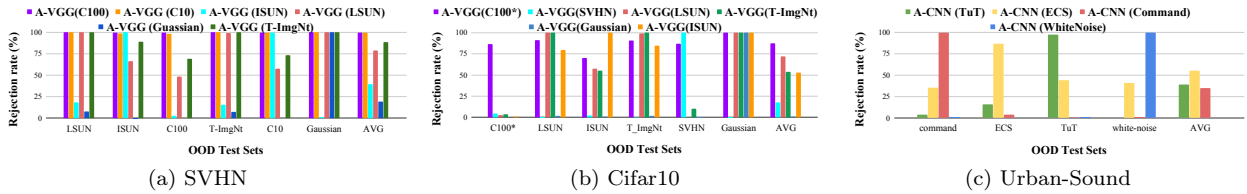[8] Instead of summation in Eq. 8, we take the average.

Figure 4: Rejection rates of A-CNNs on different OOD sets (x-axis), where each A-CNN is trained on a single OOD set, e.g. for SVHN in Figure (a) A-VGG(C100) means an Augmented VGG trained on SVHN as in distribution and CIFAR-100 as OOD set.

In Fig 4, A-CNNs trained on the most protective set (identified by our metrics) across various in-distribution tasks consistently outperform the A-CNNs trained on other OOD sets, particularly the A-CNN trained on the least protective one. For instance, A-CNN trained for CIFAR-10 with CIFAR-100* (the non-overlapped version of CIFAR-100) as the most protective OOD set has 85% rejection rate on average while the least protective one, i.e. SVHN, delivers A-CNN with the lowest average of rejection rates (21%) of OOD sets.

It is also interesting to note that even though one may expect Gaussian noise (i.e. white noise) to have well distributed samples, SE metric shows that its samples are actually not evenly distributed over all of the in-distributed sub-manifolds (having lowest SE) and sometimes (for CIFAR-10 and Urban-Sound in-distribution sets) they even have a small coverage rate. As a result, an A-CNN trained on Gaussian noise as an OOD set has the lowest average OOD rejection rate.

Consequently, the results show that all of the OOD sets are not equal for training well-generalized A-CNNs as they do not equally protect in-distribution sub-manifolds. *Thus, we highlight that protectiveness can be an important factor for differentiating OOD sets to ultimately select the most proper one.* Moreover, to select an such OOD set, we remark that our metrics are computationally inexpensive than explicitly optimizing Eq. 8, which is equivalent to searching exhaustively all A-CNNs, where each is trained on an OOD set.

In Table 1, *in-distribution* generalization performance of two A-CNNs, one trained on the most protective OOD set (named A-CNN*) and another trained on the least protective one (named A-CNN‡), are compared with their standard (Vanilla) CNN. Although the accuracy rates of the A-CNNs* drop slightly, their error rates (i.e., risks) are considerably smaller than their counterparts, i.e., vanilla CNNs. This is because the A-CNNs are able to reject some "hard" in-distribution samples, instead of incorrectly classifying them (similar to [9]). Rejecting a sample rather than incorrectly classifying it is an important aspect, particularly for security and safety concerns.

Confidence-Calibrated vanilla CNN: Instead of A-CNN, now we use calibrated CNN as the end-to-end model in order to show the different impacts of OOD sets on training well-performing calibrated CNNs. As it can be seen in Table 2, the most protective OOD set recognized by our metrics is leading to a calibrated CNN with a considerable lower average of FPR at 95% TPR and highest AUROC. While the calibrated CNN training on the least protective one has the higher FPR and the lower AUROC[9]. As a result, we highlight that efficiently recognizing proper (i.e. protective) OOD sets among the enormous available ones is a key for training a well-performed *end-to-end model* (either the underlying model is A-CNN or calibrated vanilla CNN).

In-distribution	Seen OOD set	Unseen OOD sets Avg AUROC/ Avg FPR
SVHN	‡ISUN	94.73/31.97
	LSUN	99.25/ 4.39
	C10	99.75/0.41
	T-ImgNt	99.75/1.10
	⋆C100	**99.86/0.07**
CIFAR-10	‡SVHN	86.38 /75.04
	ISUN	86.20/77.03
	LSUN	93.31/ 38.59
	T-ImgNt	**93.89**/34.44
	⋆C100*	93.03/**26.13**
Urban-Sound	‡Command	59.15/63.06
	‡TuT	45.40/85.08
	⋆ECS	**71.41/60.67**

Table 2: The effect of OOD set selection on the performance of calibrated CNNs, where each trained on an OOD set, then evaluated on unseen OOD sets. We report the average of AUROC and FPR of calibrated CNNs on unseen OOD sets and test in-distribution set.

3.2 Black-box Fast Gradient Sign (FGS) Adversaries as OOD samples

FGS adversaries with high noise level can be regarded as synthetic OOD samples, where they most likely lie out of in-distribution sub-manifolds [10, 22]. Even though such FGS adversaries contain perceptible noise, they can still fool vanilla CNNs easily [10, 30]. To explore the capability of A-CNN* in detecting such non-optimal adversaries, A-CNN*, A-CNN‡, and their vanilla counterparts are compared w.r.t. their error rates on FGS adversaries with a varying amount of noise. We generated 5,000 black-box FGS adversaries (from training in-distribution set) using another pre-trained vanilla CNN

In-dist. task	Network	In-distribution Acc (↑) / Rej (↓) / Err (↓)	OOD sets Avg OOD Rej. (↑)
SVHN	Vanilla VGG	**95.53** / − / 4.47	−
	A-VGG‡ (ISUN)	95.11 / 0 / 4.89	47.23
	A-VGG* (C100)	95.38 / 0.34 / **4.28**	**99.88**
CIFAR-10	Vanilla VGG	**88.04** / − / 11.95	−
	A-VGG‡ (SVHN)	87.75 / 0.03 / 12.22	21.41
	A-VGG* (C100*)	85.37 / 5.65 / **8.97**	**85.10**
Urban-Sound	Vanilla CNN	**67.27** / − / 32.73	−
	A-CNN‡ (Command)	65.05 / 2.02 / 32.93	26.07
	A-CNN* (ECS)	63.13 / 12.02 / **24.85**	**55.40**

Table 1: The influence of selected most and least protective OOD sets on inducing well-generalized A-CNNs with high OOD detection rates.

---

[9] For brevity, we report the average of AUROCs and FPRs of *unseen* OOD sets.

(a) Rejection rate of SVHN adversaries

(b) CD of SVHN adversaries

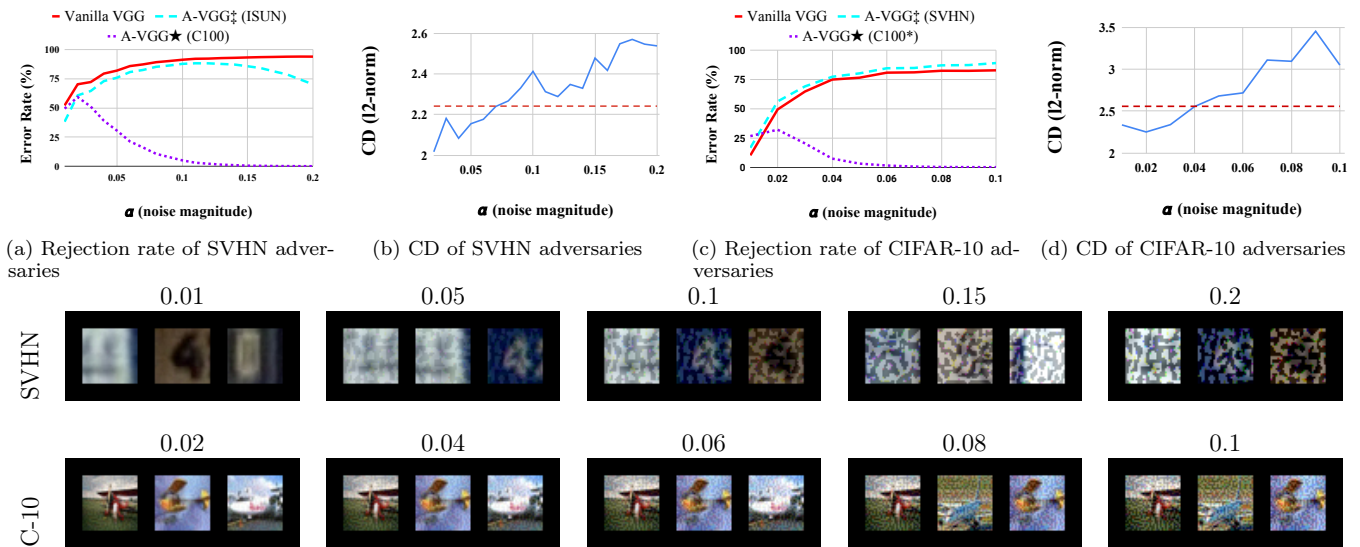(c) Rejection rate of CIFAR-10 adversaries

(d) CD of CIFAR-10 adversaries

Figure 5: FGS adversaries with various noise magnitude (shown in the two last rows). Sub-figures (a,c) show error rates of vanilla CNN, A-CNN*, A-CNN$\ddagger$ on FGS adversaries with varying noise for SVHN and CIFAR-10, respectively. Note Err rate = 1-(Acc rate +Rej rate). (b,d) Coverage Distance (CD) of FGS adversaries (their average distance to in-distribution sub-manifolds) for SVHN and CIFAR-10 respectively. The dotted red line is the Coverage Distance of the most protective OOD set, which is used to train A-CNN*.

(different from the one evaluated here). Some adversarial examples with various amounts of noise (i.e. α) are displayed in Fig 5.

As evident from Fig 5, the error rates (i.e., 1-Acc) of vanilla CNNs increase as α becomes larger, showing the transferability of these black-box FGS adversaries. In contrast, the error rates (i.e., 1-(Acc+Rej)) of the A-CNNs* approach zero as α increases since many of these FGS samples are rejected by A-CNNs*. On the contrary, the error rates of A-CNN$\ddagger$ are almost as high as those of vanilla CNNs for FGS adversaries with different magnitudes of noise. Fig 5 (b) and (d) can explain this phenomenon; larger α causes generated FGS adversaries to be further away from the sub-manifolds (i.e., larger CD). When FGS adversaries enter the protected regions by A-CNN* (starting at the distance denoted by CD of the most protective OOD set, i.e., dotted red horizontal line), they are automatically rejected as OOD samples.

4 Related Work

In [14], the authors have demonstrated that OOD samples can be discriminated from in-distribution ones by their predictive confidence scores provided by vanilla CNNs. As this baseline does not create a significant detection rate, many researchers [21, 20, 17, 2, 7, 18] have attempted to process the confidence score for creating a larger gap between in-distribution and OOD samples.

Other researchers have proposed to train end-to-end calibrated networks for making low confidence prediction on OOD sets while keeping in-distribution performance high. For example, [23] have incorporated an OOD set to in-distribution set to train a modified Siamese network in order to keep in-distribution samples nearby while pushing OOD training samples away from in-distribution ones. Others [15, 19, 31] have

proposed to train a vanilla CNN on OOD set along with in-distribution set to force explicitly prediction of OOD samples with uncertainty while confidently and correctly classifying in-distribution samples. To train such end-to-end CNN-based models, one can leverage a natural OOD set likewise [4, 15, 23] or a set of synthetically-generated OOD samples [19, 33, 13]. Apart from computational cost of generating such a set of synthetic samples, Hendrycks et al.[15] have shown a calibrated CNN trained on a proper naturalistic OOD set can outperform that of trained on GAN-generated synthetic samples.

5 Conclusion

Our main goal in this paper is to characterizing properties of OOD sets for recognizing a proper one for training an end-to-end A-CNN and calibrated vanilla CNN with high detection rate on unseen OOD sets while maintaining in-distribution generalization performance. To this end, we feature an OOD set as proper if it can cover all of the in-distribution's sub-manifolds in the feature space (i.e. protective OOD set). Then, we propose computationally efficient metrics as a tool for differentiating OOD sets for the purpose of selecting the most protective one. Finally, we empirically exhibit training end-to-end models on the most protective OOD set leads to remarkably higher detection rates of unseen OOD sets, in comparison with those models trained on the least protective OOD set. A Growing number of available OOD sets is a possible rich source for training well-performing end-to-end models to tackling OOD detection challenge, if the most proper OOD set (equivalently the most protective one) can be efficiently recognized.

REFERENCES

[1] Mahdieh Abbasi, Arezoo Rajabi, Christian Gagné, and Rakesh B Bobba, 'Towards dependable deep convolutional neural networks (cnns) with out-distribution learning', *Dependable and Secure Machine Learning*, (2018).

[2] Abhijit Bendale and Terrance E Boult, 'Towards open set deep networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, (2016).

[3] Yoshua Bengio, 'Deep learning of representations: Looking forward', in *Statistical language and speech processing*, 1–37, Springer, (2013).

[4] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić, 'Discriminative out-of-distribution detection for semantic segmentation', *arXiv preprint arXiv:1808.07703*, (2018).

[5] Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She, 'Why deep learning works: A manifold disentanglement perspective', *IEEE transactions on neural networks and learning systems*, **27**(10), 1997–2008, (2015).

[6] Qing Da, Yang Yu, and Zhi-Hua Zhou, 'Learning with augmented class by exploiting unlabeled data', in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, (2014).

[7] Terrance DeVries and Graham W Taylor, 'Learning confidence for out-of-distribution detection in neural networks', *arXiv preprint arXiv:1802.04865*, (2018).

[8] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner, 'Detecting adversarial samples from artifacts', *arXiv preprint arXiv:1703.00410*, (2017).

[9] Yonatan Geifman and Ran El-Yaniv, 'Selectivenet: A deep neural network with an integrated reject option', *International Conference on Machine Learning (ICML)*, (2019).

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and harnessing adversarial examples', *arXiv preprint arXiv:1412.6572*, (2014).

[11] Manuel Gunther, Steve Cruz, Ethan M Rudd, and Terrance E Boult, 'Toward open-set face recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 71–80, (2017).

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, 'On calibration of modern neural networks', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, (2017).

[13] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf, 'Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–50, (2019).

[14] Dan Hendrycks and Kevin Gimpel, 'A baseline for detecting misclassified and out-of-distribution examples in neural networks', *arXiv preprint arXiv:1610.02136*, (2016).

[15] Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich, 'Deep anomaly detection with outlier exposure', *Internation Conference on Representation Learning (ICLR)*, (2019).

[16] Fu Jie Huang and Yann LeCun, 'Large-scale learning with svm and convolutional for generic object categorization', in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pp. 284–291. IEEE, (2006).

[17] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta, 'To trust or not to trust a classifier', in *Advances in neural information processing systems*, pp. 5541–5552, (2018).

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, 'Simple and scalable predictive uncertainty estimation using deep ensembles', in *Advances in Neural Information Processing Systems*, pp. 6405–6416, (2017).

[19] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin, 'Training confidence-calibrated classifiers for detecting out-of-distribution samples', *International Conference on Learning Representations (ICLR)*, (2017).

[20] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, 'A simple unified framework for detecting out-of-distribution samples and adversarial attacks', in *Advances in Neural Information Processing Systems*, pp. 7167–7177, (2018).

[21] Shiyu Liang, Yixuan Li, and R Srikant, 'Principled detection of out-of-distribution examples in neural networks', *International Conference on Learning Representation*, (2018).

[22] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey, 'Characterizing adversarial subspaces using local intrinsic dimensionality', *International Conference on Learning Representations (ICLR)*, (2018).

[23] Marc Masana, Idoia Ruiz, Joan Serrat, Joost van de Weijer, and Antonio M Lopez, 'Metric learning for novelty and anomaly detection', *British Machine Vission Conference*, (2018).

[24] Alexander Meinke and Matthias Hein, 'Towards neural networks that provably know when they don't know', in *International Conference on Learning Representations (ICLR)*, (2020).

[25] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, 'TUT database for acoustic scene classification and sound event detection', in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, (2016).

[26] Karol J. Piczak, 'Esc: Dataset for environmental sound classification', in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press, (2015).

[27] Justin Salamon and Juan Pablo Bello, 'Deep convolutional neural networks and data augmentation for environmental sound classification', *IEEE Signal Processing Letters*, **24**(3), 279–283, (2017).

[28] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, 'A dataset and taxonomy for urban sound research', in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044. ACM, (2014).

[29] Shai Shalev-Shwartz and Shai Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge university press, 2014.

[30] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel, 'The space of transferable adversarial examples', *arXiv preprint arXiv:1704.03453*, (2017).

[31] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke, 'Out-of-distribution detection using an ensemble of self supervised leave-out classifiers', in *The European Conference on Computer Vision (ECCV)*, (September 2018).

[32] Pete Warden, 'Speech commands: A dataset for limited-vocabulary speech recognition', *arXiv preprint arXiv:1804.03209*, (2018).

[33] Yang Yu, Wei-Yang Qu, Nan Li, and Zimin Guo, 'Open-category classification by adversarial sample generation', *International Joint Conference on Artificial Intelligence (IJCAI)*, (2017).