In- or Out-of-Distribution Detection via Dual Divergence Estimation

Sahil Garg[1,*] **Sanghamitra Dutta**[2] **Mina Dalirrooyfard**[1] **Anderson Schneider**[1] **Yuriy Nevmyvaka**[1]

[1]Dept. of Machine Learning Research, Morgan Stanley, New York, New York, USA
[2]Dept. of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, USA
[*]Corresponding Author: sahil.garg@morganstanley.com, sahil.garg.cs@gmail.com

Abstract

Detecting *out*-of-distribution (OOD) samples is a problem of practical importance for a reliable use of deep neural networks (DNNs) in production settings. The corollary to this problem is the detection *in*-distribution (ID) samples, which is applicable to domain adaptation scenarios for augmenting a train set with ID samples from other data sets, or to continual learning for replay from the past. For both ID or OOD detection, we propose a principled yet simple approach of (empirically) estimating KL-Divergence, in its *dual form*, for a given test set w.r.t. a known set of ID samples in order to quantify the contribution of each test sample individually towards the divergence measure and accordingly detect it as OOD or ID. Our approach is compute-efficient and enjoys strong theoretical guarantees. For WideResnet101 and ViT-L-16, by considering ImageNet-1k dataset as the ID benchmark, we evaluate the proposed OOD detector on 51 test (OOD) datasets, and observe drastically and consistently lower false positive rates w.r.t. all the competitive methods. Moreover, the proposed ID detector is evaluated, using ECG and stock price datasets, for the task of data augmentation in domain adaptation and continual learning settings, and we observe higher efficacy compared to relevant baselines.

1 INTRODUCTION

Despite the great success of deep neural nets, there are important challenges that remain to be addressed in continual lifelong learning settings [Lopez-Paz and Ranzato, 2017, Riemer et al., 2018, Parisi et al., 2019, Rao et al., 2019, Lesort et al., 2021]. In continual learning settings, due to the inherent nonstationarity of a domain, it is typical to observe samples in a test setting which are Out-Of-Distribution (OOD) w.r.t. the training set. A DNN must be capable to detect such OOD samples and acknowledge that it is not knowledgeable to have high confidence outputs on such inputs Hendrycks and Gimpel [2016], Liu et al. [2020], Hendrycks et al. [2022]. In the extreme scenarios where a majority of test samples are OOD w.r.t. the training set, it is natural to retrain the network on observations from the most recent past as representative of the (OOD) test setting. In such case of continual learning or domain adaption, for data augmentation, one can detect all the samples from the history of the same domain or even other domains which are ID w.r.t. the representative set. Both the problems of ID or OOD detection are related in theory, two sides of the same coin, yet differing in their utilities for lifelong learning. We

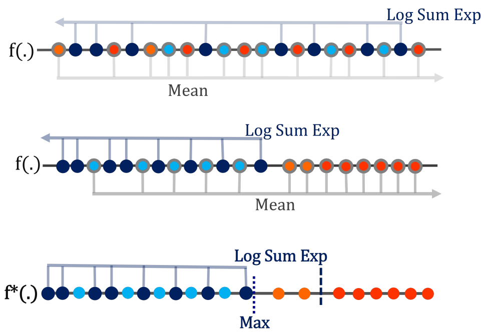


Figure 1: The dark blue dots are ID samples and all the other dots (with gray circles) are test samples. For detecting OOD samples in a test set, we propose to first empirically estimate KL-Divergence (KL-D), in its dual form, between the underlying distribution of test samples and of the ID samples. For estimating the KL-D in its dual form, ID and test samples are mapped to 1-D space by a dual function, $f(.)$, and the dual space is optimized by minimizing log sum exp *smooth max*) of dark blue dots (ID samples) while maximizing mean of rest of the dots (test samples). In the second sketch, where the dual space is nearly-optimized, ID (light blue) and OOD (red and orange) samples within the test set are well separated.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

approach this problem from the perspective of estimating divergence between two distributions. In a continual learning setting, if one encounters a test set of samples which are all known to be OOD w.r.t. the known set of ID samples, we should expect high divergence between the underlying distributions of the two sets. In contrast, if a test contains only ID samples, divergence between the respective distributions should be close to zero. Of course, in the real world settings, a test set may contain a mixture of both ID and OOD samples. In such case, we wish to split the test set into two parts, OOD vs ID samples. The aforementioned intuitions about pure ID vs pure OOD set should apply to the two parts of the test set as well.

Specifically, we propose a *novel* approach for ID vs OOD detection based upon the concept of dual divergence estimation. As illustrated in Fig. 1, we propose to empirically estimate KL-Divergence of a given test set of samples w.r.t. a set of known ID samples. The key idea is that by estimating KL-Divergence in its dual form due to Donsker and Varadhan [1975], we obtain the individual contribution of each sample in the test set towards the divergence measure so as to detect OOD and ID samples within the test set. Our principled approach enables OOD detection algorithms that enjoy linear time complexity and *theoretical guaranties*.

For the problem of OOD detection in pre-trained deep neural nets, Imagenet has been the most challenging and well known benchmarked ID dataset. We specifically consider the task of OOD detection in a WideResnet and a Vision Transformer (ViT) pretrained on Imagenet. For OOD datasets, we consider 51 datasets from diverse domains including all four previously benchmarked OOD datasets. Our extensive empirical analysis shows that the proposed OOD detector is consistently and drastically superior w.r.t. all the competitive methods.

For the evaluation of ID detector, we consider the problem of timeseries forecasting. We augment the training dataset for a given timeseries with ID samples detected from the past of the same timeseries as well as of other timeseries. Our empirical analysis of data sets of US stock prices and ECG demonstrates the competitiveness of our proposed ID detector w.r.t. relevant baselines of contextual replay from continual learning and of domain adaptation.

Contributions Our contributions are: (i) a novel principled information theoretic approach for OOD detection which enjoys theoretical guaranties; (ii) extensive empirical evaluation demonstrating the superiority of our approach while also establishing new benchmarks on 47 new OOD datasets considering Imagenet-1k as the ID dataset; (iii) we leverage the proposed approach for ID detection to augment data sets in continual learning or domain adaptation settings as demonstrated using multivariate timeseries datasets from the diverse domains of finance and healthcare. (iv) Codebase at github.com/morganstanley/MSML/tree/main/papers.

1.1 RELATED WORKS FOR OOD DETECTION

Detecting OOD samples in a pre-trained deep neural network (DNN) is a problem of high practical importance. The intuition behind this body of work is that the representations from the top hidden layer of a DNN (referred to as logits) are informative of all the hierarchical features relevant for distinguishing between ID classes as well as for differentiating OOD vs ID samples. Various heuristics have been proposed in regards to what kind of information in the logits may be relevant for detecting OOD samples. Hendrycks and Gimpel [2016] propose maximum softmax probability (MSP) as the criterion for OOD detection. However, in practice, it is observed that DNNs tend to assign a high probability to one of the ID classes even for OOD samples. Liang et al. [2018] propose ODIN which attempts to fix this issue by temperature scaling and adding small perturbations to the inputs.

Liu et al. [2020] obtain energy score from the logits as the criterion for OOD detection. Intuitively, since the energy score (EBO) is $\log$ *sum* $\exp$ function of the logits which is a well known approximation (upper bound) for $\max$ function, it accounts for all the high values in the vector of logits rather than relying upon the highest value only, effectively a smooth version of $\max$ function. For the same reason, it is known to be consistently superior to ODIN and its predecessors, as we also observe in our extensive evaluation. Hendrycks et al. [2022] propose to simply use the maximum value of the logits as the criterion for OOD detection. Their intuition about the superiority of raw logits over its normalization as in MSP or ODIN is shown to be empirically valid. Though their approach is not as effective as energy scores ("smooth" max of the logits) for the reasons mentioned above. Hendrycks et al. [2022] also introduce KL-Matching method which computes relative entropies of per class of a softmax distribution w.r.t. the respective distribution templates.

Huang et al. [2021] propose to use the gradient norm of KL-Divergence of the softmax distribution w.r.t. its respective uniform distribution, relying upon the intuition that the gradients should be of higher norm for ID samples in contrast to OOD samples. There isn't solid evidence supporting this intuition, it may be applicable only to certain OOD scenarios. Sun et al. [2021] propose a simple yet highly effective technique of rectifying activations, i.e. truncating activations above a certain threshold, in the penultimate layer of a DNN. The authors suggest that ReAct is particularly suitable when OOD activations are chaotic and positively skewed in comparison to ID activations.

Wang et al. [2022b] propose to robustify existing OOD detection methods by watermarking ID patterns through reprogramming of the neural nets. Specifically, a static pattern is learned as a watermark to be added to any given input for its detection as OOD. Another technique for robustifying

OOD methods is to sparsify *weights* at inference time [Sun and Li, 2022]. Djurisic et al. [2022] instead propose to sparsify *representations* in the top layers of a DNN which is superior to sparsifying weights as per our analysis.

Another body of work is based on non-parametric modeling of the logits, such as deep Gaussian Mixtures [Morteza and Li, 2022], or deep k Nearest Neighbors [Sun et al., 2022]. Gomes et al. [2022] propose to employ Fisher-Rao distances between normalized logits (softmax distribution) for obtaining centroids as representative of ID classes. These nonparametric methods, as also acknowledged by Sun et al. [2022], are inefficient if the number of ID classes is large (1000) as in Imagenet dataset.

Our approach is generic enough to be applicable to other various flavors of OOD detection problem settings such as retraining strategies or regularization techniques, advanced techniques for generating OOD samples, mixing raw features and representations from different layers, generative modeling or autoencoding, out-of-model-scope detection, etc. [Ren et al., 2019, Teney et al., 2020, Bitterwolf et al., 2020, Mahmood et al., 2020, Morningstar et al., 2021, Fort et al., 2021, Zhou et al., 2021, Ming et al., 2022a,b,c, Du et al., 2022, Yang et al., 2022, Lin et al., 2022, Wang et al., 2022a, Wei et al., 2022, Liu et al., 2022, Fan et al., 2022, Jiang et al., 2022, Guérin et al., 2022, Wu et al., 2022, Wang et al., 2022c, Huang et al., 2022, Zhang et al., 2022, Wilson et al., 2023, Wang et al., 2023], though we restrict our analysis in this paper to the above discussed settings of OOD detection in pretrained DNNs for its high practical importance and simplicity.

1.2 RELATED WORKS FOR ID DETECTION

In a general continual learning (CL) scenario, the goal is to ensure that a neural net does not catastrophically forget what it has learned in the past when it learns from the present episodes Riemer et al. [2018], McCloskey and Cohen [1989]. One of the most simple, effective, brain-inspired, and generic non-intrusive approach for continual learning is *replay* of the past memory episodes [Rolnick et al., 2019, van de Ven et al., 2020, Deja et al., 2021].

Shin et al. [2017] introduced the idea of deep generative replay, popularly known as DGR. To alleviate the problem of catastrophic forgetting, [Van de Ven and Tolias, 2018] propose generative replay via distillation (i.e., employing class probabilities as “soft targets”). [Aljundi et al., 2019] address the issue of forgetting by formulating a controlled sampling criterion for both generative and experience replay settings. Buzzega et al. [2020] propose “dark experience replay" as a simple yet powerful baseline that mixes rehearsal with knowledge distillation and regularization. The problem of contextual replay also has connections to the classical field of active learning, and it has been pursued in the recent works of continual learning as well [Tang and Matteson, 2020, Sun et al., 2021]. Sun et al. [2021] propose to quantify informativeness of samples via criteria of surprise and learnability.

The CL setting considered in this paper is related to the above but different in the sense that we are interested in leveraging all the knowledge (observations) from the past of a given domain or of other domains for training a model representative of the present only. While we are interested only in detecting ID samples and not generating novel ones or adapting existing ones, domain adaptation techniques such as optimal transport or domain discrimination are also tangentially related [Ganin et al., 2016, Balaji et al., 2020].

2 DDE FOR OOD OR ID DETECTION

We discuss our novel approach for OOD detection based on dual divergence estimation, and its applicability for ID detection in continual learning settings.

Problem Settings For the problem of OOD detection, we assume the availability of an ID set of N samples, $\mathbf{X}^{in} = \{\mathbf{x}_i^{in}\}_{i=1}^{N}$. For OOD detection in pretrained networks, $\mathbf{X}^{in}$ is assumed to be a set of the representations of ID inputs from the top hidden layer of a DNN, also referred as logits. Given a test set of inputs or the respective representations, $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{M}$, we detect if $\mathbf{x}_j$ is out-of-distribution w.r.t. the underlying (unknown) distribution $\mathcal{X}^{in}$ of $\mathbf{X}^{in}$.

In our approach for OOD detection, we make no assumptions about the underlying distribution, nor do we advocate learning the corresponding density function. Moreover, our approach is not to learn an OOD detector prior to encountering a test set, but rather to estimate *empirical* KL-Divergence (KL-D) between a test set and the ID set on the fly. The key idea is that estimating the divergence measure in its *dual form* is naturally informative about the subset of samples in the test set that are OOD.

Mathematically, KL-D between the (unknown) underlying distribution of a test set $\mathcal{X}$ and the (unknown) representative distribution $\mathcal{X}^{in}$ is expressible as,

$$D(\mathcal{X} \| \mathcal{X}^{in}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{X}} \log \frac{P(\mathbf{x})}{P^{in}(\mathbf{x})} \tag{1}$$

Here, $P(.)$ and $P^{in}(.)$ are density functions corresponding to the distributions $\mathcal{X}$ and $\mathcal{X}^{in}$. As such, even for the empirical estimate of KL-D in Eq. 1, one has to rely upon the knowledge of the density functions which are unknown. Estimating a density function is a hard problem on its own and can be avoided if estimating the divergence in its dual form by Donsker and Varadhan [1975], as expressed below.

$$D(\mathcal{X} \| \mathcal{X}^{in}) = \max_{f(.)} \mathbb{E}_{\mathbf{x} \sim \mathcal{X}} f(\mathbf{x}) - \log \mathbb{E}_{\mathbf{x}^{in} \sim \mathcal{X}^{in}} e^{f(\mathbf{x}^{in})}$$

Herein, $f(.)$ can be any function such that the expectations are finite, referred as the *dual function*. As we observe above, for estimating KL-D in its dual form, we only need samples from the distributions $\mathcal{X}$ and $\mathcal{X}^{in}$, not the density functions. This form is particularly suitable for an *empirical* estimate of KL-D between $\mathcal{X}$ and $\mathcal{X}^{in}$ using a test set $\mathbf{X}$ and the ID set $\mathbf{X}^{in}$ as below.

$$\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max_{\hat{f}(.)\in\mathcal{H}} \sum_{\mathbf{x}_j\in\mathbf{X}} \frac{\hat{f}(\mathbf{x}_j)}{M} - \log \sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} \frac{e^{\hat{f}(\mathbf{x}_i^{in})}}{N}$$

Here, the maximization is performed over the fixed class of functions $\mathcal{H}$, e.g., the class of functions that can be learned by DNNs [Belghazi et al., 2018]. As we also illustrate in Fig. 1, all the samples from both sets are mapped into the 1-D dual functional space that is optimized such that expected value of the test samples is maximized whereas the $\log$ sum $\exp$ i.e. smooth max of the ID samples is minimized.

In the optimal (1-D) dual space (f^*) as shown in Fig. 2, test samples (light blue, orange, and red dots) and ID samples (dark blue dots) are separated as much as possible. The red dots from the test set are located on the r.h.s. in the optimal dual space contributing to the KL-D measure. On the other hand, despite the objective of maximizing the value of light blue dots (test samples) while minimizing smooth max of the ID samples, we find the light blue dots interspersed between the dark blue ones (ID samples), clearly separated from red dots. This is simply because the optimized dual function fails to distinguish such samples from the ID set. Therefore, intuitively, these light blue test samples should be detected as ID samples whereas the red dots can be detected as OOD samples. Later, we support these intuitions with theoretical analysis. The optimized dual space provides us a nice geometric interpretation of ID vs OOD samples. On this note, in Fig. 2, it is also interesting to observe that orange dots (test samples) lie on the soft boundary between the detected ID and OOD samples in the test set.

In essence, estimating KL-D of a test set w.r.t. the ID set in its dual form naturally splits it into two subsets, detected ID (light blue dots) vs detected OOD samples (red dots). Our approach has the advantage of treating OOD detection problem purely as an optimization problem (optionally solvable using deep learning as we propose) while enjoying information theoretic guaranties.

Although, in practice, one can choose any point in the optimized (1-D) dual space as a cut point (threshold) for OOD detection, we approach this problem more formally. Intuitively, the smooth max of the dark blue dots could serve as a cut point for OOD detection as we also illustrate in Fig. 2. In the following, we present some theoretical insights which confirm this intuition. First, we establish that for detecting OOD samples in a test set, there is no cut point required on the left side of ID samples in the optimized (1-D) dual space, i.e. no test samples exist beyond the left boundary of ID samples (dark blue dots) in Fig. 2.

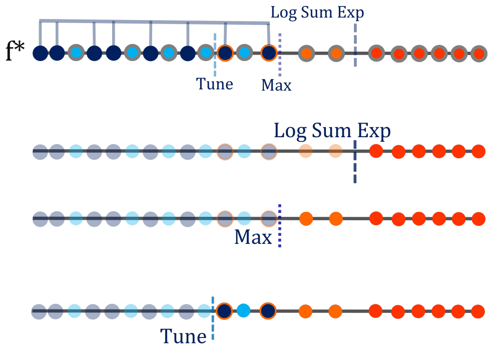


Figure 2: Our three choices for cut points in the optimized 1-D dual functional space for OOD detection. The dark blue dots refer to samples from the ID set whereas the rest of the dots are points from a given test set. From a theoretical standpoint, we propose to compute the cut point as smooth max (*Log Sum Exp*) of the dark blue dots. One interpretable choice of cut point for OOD detection is the maximum of the dark blue dots (*Max*); this cut point detects test points on the boundary of the ID set as OOD as well (orange dots). In a scenario where the ID set is corrupted by OOD samples (for example, see the dark blue dot with orange circle), it would make sense to find a cut point by tuning within the range of dark blue dots (*Tune*), thus deeming even some of the known ID samples as OOD.

Theorem 1 *Given an ID set $\mathbf{X}^{in}$ and a test set $\mathbf{X}$, from estimating KL-D of $\mathbf{X}$ w.r.t. $\mathbf{X}^{in}$ in its dual form as,*

$$\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max_{\hat{f}(.)\in\mathcal{H}} \sum_{\mathbf{x}_j\in\mathbf{X}} \frac{\hat{f}(\mathbf{x}_j)}{M} - \log \sum_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} \frac{e^{\hat{f}(\mathbf{x}_i^{in})}}{N},$$

we obtain the optimal dual function $\hat{f}^*(.)$. The optimal dual function $\hat{f}^*(.)$ satisfies the following:

$$\forall \mathbf{x}_j \in \mathbf{X}, \quad \min_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} \hat{f}^*(\mathbf{x}_i^{in}) \le \hat{f}^*(\mathbf{x}_j). \tag{2}$$

Our next result establishes that OOD samples if any lie only on the right side of the ID samples (dark blue dots) in the optimized (1-D) dual space.

Theorem 2 *Given an ID set $\mathbf{X}^{in}$ and a test set $\mathbf{X}$, the optimal dual function $\hat{f}^*(.)$ which maximizes the estimate of KL-D as defined in Theorem 1, satisfies the following:*

$$\max_{\mathbf{x}_i^{in}\in\mathbf{X}^{in}} \hat{f}^*(\mathbf{x}_i^{in}) \le \max_{\mathbf{x}_j\in\mathbf{X}} \hat{f}^*(\mathbf{x}_j) \tag{3}$$

These results provide critical intuition on the role of the optimal dual function $\hat{f}^*$ in distinguishing ID and OOD samples. As discussed above, essentially, the function $\hat{f}^*$ attempts to find a representation where the ID and OOD samples are maximally separated. So, points that lie in $\mathbf{X}^{in}$ are assigned a lower value, and points in $\mathbf{X}$ are assigned values based on how similar they are to $\mathbf{X}^{in}$. Thus, test ID

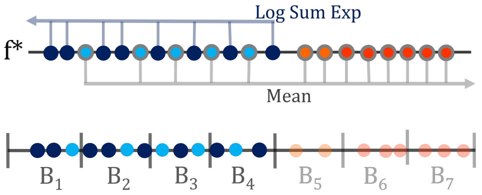


Figure 3: ID sampling from the 1-D dual space. The dark blue dots refer to samples from the ID set whereas the rest of the dots are points from a given test set. All the data points are binned using histograms of constant width d in the optimal (1-D) dual space. We propose to take test samples from bins B_1, B_2, B_3, B_4 as ID samples, the ones with light blue color.

samples are assigned somewhat lower values as well and test OOD samples are assigned higher values. In Theorem 2, a strict inequality holds when there exists a cut point such that some points $\mathbf{P} \subseteq \mathbf{X}$ that are possibly quite dissimilar to $\mathbf{X}^{in}$ can be separated from $\mathbf{X}^{in} \cup (\mathbf{X} \backslash \mathbf{P})$ using $\hat{f}^*$. When equality holds, it can be interpreted as: $\mathbf{X}$ and $\mathbf{X}^{in}$ are already quite similar. As per the intuition mentioned above, we *propose to use the smooth max of the ID samples (dark blue dots) as a cut point for OOD detection*, with theoretical guarantees as presented below in Theorem 3.

Theorem 3 *Given an ID set $\mathbf{X}^{in}$ and a test set $\mathbf{X}$, let $\hat{f}^*(.)$ be the optimal dual function which maximizes the estimate of KL-D (as defined in Theorem 1). Then, for the subset of the test set deemed as OOD,*

$$\mathbf{X}^{ood} = \{\mathbf{x}_j : \mathbf{x}_j \in \mathbf{X}, \hat{f}^*(\mathbf{x}_j) > \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \exp \hat{f}(\mathbf{x}_i^{in})\},$$

its KL-D w.r.t. the ID set is lower bounded as,

$$\hat{D}(\mathbf{X}^{ood} \| \mathbf{X}^{in}) > \log(N).$$

The above result suggest that OOD detection in DNNs can be improved by increasing the sample size in training (ID) sets. Another way to interpret the lower bound is in its relation to the entropy of the ID set. Since all the empirical realizations in the ID set are equally probable, as such, we obtain the maximum possible value of the entropy as the lower bound. Furthermore, this result is applicable for any subset of the detected OOD samples as well. One alternative, as shown in Fig. 2, is to use $\max$ instead of the smooth max for finding the cut point. Having a cut point any further left of the above choices can be problematic, though relevant if the ID set consists of OOD samples as noise.

See the supplement for details on how to employ DNNs as the dual function approximators.

2.1 ID DETECTION

Intuitively, the problem of ID detection is similar to the problem of detecting OOD samples though its potential use cases are different. For this problem, $\mathbf{X}^{in}$ denotes the representative set of samples such as observations from the present (recent past) of a domain of interest. On the other hand, $\mathbf{X}$ refers to a large set of observations from the historical past of the same domain or other related domains.

Similar to OOD detection, we estimate divergence of $\mathbf{X}$ w.r.t. $\mathbf{X}^{in}$ in its dual form. The key difference being that we perform histogram binning of all the samples from both the sets in the 1-D dual functional space [Freedman and Diaconis, 1981]. The bins which contain samples only from $\mathbf{X}$ and not from $\mathbf{X}^{in}$ are discarded, and we select all the samples of $\mathbf{X}$ from the rest of the bins as ID detections w.r.t. $\mathbf{X}^{in}$. We also introduce the following notation for the evaluation of the divergence estimate at any function $f(\cdot) \in \mathcal{H}$:

$$\hat{D}_f(\mathbf{X}_a \| \mathbf{X}_b) := \sum_{\mathbf{x}_j \in \mathbf{X}_a} \frac{f(\mathbf{x}_j)}{|\mathbf{X}_a|} - \log \sum_{\mathbf{x}_i \in \mathbf{X}_b} \frac{e^{f(\mathbf{x}_i)}}{|\mathbf{X}_b|}.$$

Theorem 4 *Given a representative ID set $\mathbf{X}^{in}$ and a set of observations from the historical past $\mathbf{X}$, from estimating KL-D of $\mathbf{X}$ w.r.t. $\mathbf{X}^{in}$ in its dual form as,*

$$\hat{D}(\mathbf{X} \| \mathbf{X}^{in}) = \max_{\hat{f}(.) \in \mathcal{H}} \sum_{\mathbf{x}_j \in \mathbf{X}} \frac{\hat{f}(\mathbf{x}_j)}{M} - \log \sum_{\mathbf{x}_i^{in} \in \mathbf{X}^{in}} \frac{e^{\hat{f}(\mathbf{x}_i^{in})}}{N},$$

we obtain the optimal dual function, $\hat{f}^*(.)$. Then, from histogram binning in the 1-D dual functional space of both the sets, we select the samples of $\mathbf{X}$ with respect to the distribution of $\mathbf{X}^{in}$ in the bins, denoted as $\bar{\mathbf{X}}$. For uniform width d of histogram bins, we have:

$$\hat{D}_{\hat{f}^*}(\bar{\mathbf{X}} \| \mathbf{X}^{in}) \leq O(d). \tag{4}$$

Here $\hat{f}^*$ is the optimal dual function that maximizes the divergence estimate between $\mathbf{X}$ and $\mathbf{X}^{in}$. We show that the divergence estimate computed between the selected ID samples $\bar{\mathbf{X}}$ and $\mathbf{X}^{in}$ (evaluated using the same $\hat{f}^*$) is quite small and in fact bounded by the histogram width. We note that while we show this result for $\hat{D}_{\hat{f}^*}(\bar{\mathbf{X}} \| \mathbf{X}^{in})$ (which may in principle be different from $\hat{D}(\bar{\mathbf{X}} \| \mathbf{X}^{in})$ since the maxima may be attained for a different function other than $\hat{f}^*$), we are also able to show that under certain restrictions on the model class $\mathcal{H}$, $\hat{D}(\bar{\mathbf{X}} \| \mathbf{X}^{in}) \approx \hat{D}_{\hat{f}^*}(\bar{\mathbf{X}} \| \mathbf{X}^{in})$ (see Theorem 5).

The significance of Theorem 4 is that the divergence of the detected ID samples from $\mathbf{X}$ w.r.t. the given ID set $\mathbf{X}^{in}$ directly depends upon the expressiveness of the histogram model itself. Note that it is not desirable to reduce the upper bound to value zero by employing a very small bin width as it would lead to selecting only those samples from $\mathbf{X}$ which are highly similar to the samples in $\mathbf{X}^{in}$, an obvious case of overfitting in sampling. Whereas choosing too large a value for bin width is not also advisable.

Compute cost for dual divergence estimation and binning is linear in sample size whereas the compute complexity for sampling from the bins is constant.

Lastly, for completeness, we include another result (Theorem 5) which demonstrates that computing the KL-D estimate using the dual optimal function $\hat{f}^*$ suffices in a lot of scenarios, particularly when using subsets of $\mathbf{X}$ separable in the dual space.

Theorem 5 *Let $\mathcal{H}$ be a class of functions such that each $f \in \mathcal{H}$ satisfies $|f(x)| < \infty \; \forall \; x \in R^k$. Furthermore, if $f_1(x), f_2(x), g(x) \in \mathcal{H}$, then functions of the form: $f(x) = f_1(x)I(g(x) \geq \tau) + f_2(x)I(g(x) < \tau)$ (which are essentially derived entirely from functions in $\mathcal{H}$) also lie in $\mathcal{H}$ for any constant τ and indicator function $I(\cdot)$. Consider $\hat{f}_1 \in \mathcal{H}$ such that $\hat{D}(\mathbf{X}\|\mathbf{X}^{in}) = \max_{f \in \mathcal{H}} \hat{D}_f(\mathbf{X}\|\mathbf{X}^{in}) = \hat{D}_{\hat{f}_1}(\mathbf{X}\|\mathbf{X}^{in})$. Then, for a subset $\bar{\mathbf{X}} \subseteq \mathbf{X}$ such that $\hat{f}_1(x) > \tau$ for $x \in \mathbf{X} \backslash \bar{\mathbf{X}}$ and $\hat{f}_1(x) \leq \tau$ for $x \in \mathbf{X}^{in} \cup \bar{\mathbf{X}}$, we have $\hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \hat{D}_{\hat{f}_1}(\bar{\mathbf{X}}\|\mathbf{X}^{in})$.*

Intuitively, what Theorem 5 demonstrates is that if there is a better function $\hat{f}_0 \in \mathcal{H}$ such that $\hat{D}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) = \hat{D}_{\hat{f}_0}(\bar{\mathbf{X}}\|\mathbf{X}^{in}) > \hat{D}_{\hat{f}^*}(\bar{\mathbf{X}}\|\mathbf{X}^{in})$, then one might be able to leverage this $\hat{f}_0$ to design an even better dual function for the original divergence estimate, i.e., there would exist an $\hat{f}_0^*$ such that $\hat{D}_{\hat{f}_0^*}(\mathbf{X}\|\mathbf{X}^{in}) > \hat{D}_{\hat{f}^*}(\mathbf{X}\|\mathbf{X}^{in}) = \hat{D}(\mathbf{X}\|\mathbf{X}^{in})$ which leads to a contradiction.

3 EMPIRICAL EVALUATION

In the following, we present our empirical analysis for both the problem of OOD and ID detection.

3.1 OUT-OF-DISTRIBUTION DETECTION

Datasets & Evaluation Settings We perform extensive empirical analysis for the problem of OOD detection in deep neural networks, WideResnet101 and ViT-L-16, pretrained on Imagenet-1k. Since there are various possible scenarios for observing OOD samples, we use an extensive list of 51 image datasets which are OOD w.r.t. Imagenet, including the previously benchmarked four datasets: SUN, Places, iNaturallist (species), and Textures. All the images are rescaled to size 224x224, following the standard procedure for preprocessing as considered in the previous works. Following the procedure of simple perturbations proposed by Hendrycks et al. [2019], we generate a validation set of OOD samples from ID samples in Imagenet. While a validation OOD set is optional for many of methods for OOD detection including ours, we find it useful for reproducibility purposes (reporting the validation accuracy) and for scenarios where multiple hyperparameter configurations of a method are equivalent if one were to only consider the standard criterion of 5% ID samples being falsely identified as OOD (corresponding to the evaluation metric FPR95) for hyperparameter tuning. See more details in the supplement.

Competitive Methods We compare our proposed approach of dual divergence estimation (***DDE****) w.r.t. a comprehensive list of methods for OOD detection in pretrained DNNs (see Sec. 1.1 for more details): (i) maximum softmax probability (*MSP*); (ii) maximum logit score (*MLS*); (iii) *ODIN* (iv) energy scores (*EBO*); (v) gradient norms (*GN*); (vi) Reactivation of representations (*ReAct*); (vii) Gaussian mixtures (*GM*); (viii) k-Nearest Neighbors (*kNN*); (ix) sparsifying weights (*DICE*); (x) sparsifying representations (*ASH*); (xi) watermarking (*WM*); (xii) KL-Matching (*KL-M*); (xiii) hyperspherical embeddings (*CIDER*); (xiv) information geometric approach of computing Fisher Rao distances between softmax distributions (*IGE*). Note that *DDE** by default employs $\max$ function based cut point in the dual functional space whereas DDE-SM* refers to *smooth max* function for the cut point, as detailed in Sec. 2.

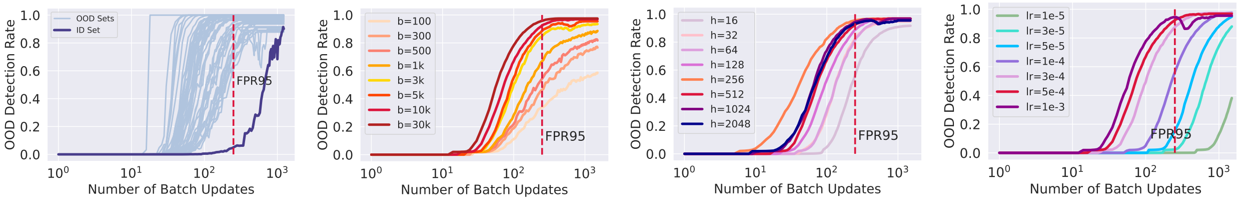
3.1.1 Empirical Results

In Table 1, we compare all the methods across 51 OOD test datasets for OOD detection in WideResnet101, using the standard and the most relevant evaluation metric, FPR95 (↓). We observe that OOD detection rate varies highly across methods and the datasets. While our methods, *DDE** and *DDE-SM**, manifests drastically lower FPR95 rates w.r.t. all the other methods, simple baselines such as *React* and *WM* perform competitively. See the supplement for our analysis on OOD detection in ViT-L-16.

Furthermore, see Fig. 5, for the analysis on detecting OOD samples at the cost of falsely detecting ID samples as OOD. Note that in our approach *DDE**, unlike the other methods, the cut point (threshold) for detecting OODs is fixed, and it gives a very low false detection rate in the ID set as such (as desired). For obtaining higher false detection rate with our method as required solely for the purpose of the analysis presented in Fig. 5, we have to overfit the neural dual function by performing a very large number of batch updates (which is not required for practical use of the detector). For the same reason, the curve for *DDE** hardly changes beyond false rate of 0.06 on X-axis. In contrast, in all the other methods, threshold for OOD detection score is manually tuned for there is a trade off between detecting OODs in ID vs OOD sets.

Dataset	**MSP**	**MLS**	**ODIN**	**EBO**	**GN**	**ReAct**	**GM**	**kNN**	**DICE**	**ASH**	**WM**	**KL-M**	**CIDER**	**IGE**	**DDE***	**DDE-SM***
ID Test ↑	93	93	92	93	95	87	**98**	95	91	93	93	94	96	92	95	94
OOD Val.	50	48	49	47	69	40	98	84	49	48	44	58	65	46	**31**	42
SUN	65	60	63	61	47	47	98	72	35	22	**12**	75	75	57	18	24
Places	68	63	66	63	61	34	97	72	47	34	57	77	77	60	**10**	21
iNaturalist	55	62	63	61	50	20	96	65	26	12	60	74	73	57	**11**	25
Textures	68	96	81	81	61	47	43	69	32	**12**	61	95	84	96	15	30
Agriculture Crop	1	**0**	1	2	82	2	100	81	9	**0**	2	16	72	**0**	**0**	**0**
Animation	42	33	40	30	94	21	100	66	38	30	29	59	69	27	**6**	19
Brain Tumors	36	26	34	21	99	16	100	69	31	20	14	54	74	20	**3**	4
Chest Xray	22	15	20	13	67	11	100	71	21	13	7	42	71	10	**4**	7
Faces in the Wild	39	29	37	26	97	19	100	68	36	26	24	57	72	23	**9**	16
Fastfood	70	64	68	62	91	47	97	70	64	60	59	79	77	60	**10**	18
Gemstone	66	59	65	54	97	39	97	63	54	52	50	77	71	52	**4**	18
LEGO	11	4	10	2	97	3	100	73	11	3	2	32	76	3	**0**	**0**
Plant Diseases	27	20	26	18	95	15	100	67	30	18	17	49	72	14	**2**	3
USPS	38	27	36	18	97	12	100	62	26	18	12	55	69	18	**1**	3
Alzeihmers	22	14	21	8	100	5	100	67	18	8	4	40	67	7	**1**	2
Blood Cells	16	11	14	13	79	13	100	72	22	10	6	37	70	9	**1**	2
Brand Logos	**0**	**0**	**0**	**0**	95	**0**	100	94	**0**	**0**	**0**	1	74	**0**	**0**	**0**
Captcha	**0**	**0**	**0**	**0**	100	**0**	100	100	**0**	**0**	**0**	**0**	100	**0**	**0**	**0**
Cards	77	74	76	73	88	59	86	67	71	70	67	83	78	71	**11**	14
Arabic Handwritten Char.	34	25	33	15	66	10	99	55	18	15	**4**	47	64	17	**4**	6
Chess Pieces	26	16	24	12	95	12	100	69	23	10	9	49	77	10	**1**	2
Chinese Fine Art	5	2	4	4	89	6	100	79	15	**1**	3	28	77	**1**	**1**	**1**
Coffee Beans	26	17	25	11	99	10	100	65	22	11	10	45	70	11	**1**	2
Colonoscopy	4	**1**	3	2	97	2	100	71	7	**1**	**1**	23	67	**1**	**1**	2
Covid CT Scans	26	18	24	14	95	11	100	69	25	15	11	48	70	12	**3**	**3**
Diamonds	47	40	45	39	97	31	100	76	46	39	36	62	78	35	**3**	5
Emotional Faces	34	25	32	21	87	15	100	68	30	21	16	52	71	18	**5**	16
Human Eyes	39	31	37	27	97	20	100	66	35	26	24	57	69	24	**5**	9
Fire & Smoke	**0**	**0**	**0**	**0**	81	**0**	100	91	**0**	**0**	**0**	**0**	72	**0**	**0**	**0**
English Handwritten Char.	26	18	25	10	69	8	99	55	16	9	9	39	60	11	**2**	3
Excavation	3	1	2	1	99	1	100	79	6	**0**	**0**	17	68	1	**0**	**0**
Eyes	33	25	32	24	88	19	100	71	31	22	11	52	72	20	**3**	4
Handwritten Math Symbols	34	24	33	15	74	10	99	53	19	13	11	48	62	15	**1**	2
Bart and Homer	**0**	**0**	**0**	**0**	100	**0**	100	78	1	**0**	**0**	11	72	**0**	**0**	**0**
Indian Food	67	62	65	61	93	49	98	68	64	56	56	79	76	58	**13**	27
LEGO Minifigures	8	4	8	2	97	3	100	74	12	2	2	26	72	1	**0**	**0**
Licence Plates	**0**	**0**	**0**	**0**	100	**0**	100	94	**0**	**0**	**0**	2	68	**0**	**0**	**0**
Meat Quality	1	**0**	1	**0**	98	**0**	100	78	4	1	**0**	10	57	**0**	**0**	**0**
Monkeypox	67	64	65	64	77	50	96	65	65	62	52	77	74	61	**8**	12
Movie Posters	57	51	55	49	94	37	100	69	55	47	48	71	76	45	**14**	24
Ornamental Plants	20	13	18	14	84	14	100	75	26	12	14	40	69	10	**0**	1
Paintings	6	3	6	5	77	5	100	69	9	3	3	28	66	2	**1**	4
Pollen Grain	25	17	23	16	94	16	100	68	30	13	12	50	73	13	**1**	**1**
QR Codes	22	13	20	9	98	7	100	71	20	10	5	41	71	8	**1**	2
Railway Tracks	2	1	2	1	82	2	100	74	6	**0**	1	20	68	1	1	1
Weed Crops	42	34	40	32	94	26	100	72	40	31	26	58	69	28	**4**	**4**
YouTube Thumbnails	54	47	52	47	91	40	100	76	54	46	44	70	80	43	**5**	19
Weather	75	72	73	73	91	58	95	78	72	73	66	80	80	70	**14**	36
Sign Language	30	20	29	13	100	10	100	62	23	12	11	48	65	13	**1**	2
Stairs	**0**	**0**	**0**	**0**	69	**0**	100	88	**0**	**0**	**0**	1	64	**0**	**0**	**0**
Shells or Pebbles	77	74	75	74	83	59	91	69	72	71	71	83	76	71	**22**	33
Summary Statistics	32±25	27±25	31±25	25±25	87±13	18±18	98±8	72±9	28±21	20±21	20±22	46±25	72±7	23±25	**4±5**	8±10

Table 1: Evaluation results for OOD detection in WideResnet101 pretrained on Imagenet-1k using the metric FPR95 (↓). Best scores are shown in bold and the second best scores are underlined.



(a) Test sets vs ID set (b) Vary Batch Size (c) Vary No. of Hidden Units (d) Vary Learning Rate

Figure 4: Ablation study for WideResnet101. OOD detection rate w.r.t. the number of batch updates performed for estimating KL-D is analyzed. Variation of OOD detection across all the test sets, and w.r.t. change in batch size, number of hidden units, learning rate is presented. FPR95 shown in each of the plots are for the default configuration only (b=10k, h=512, lr=5e-4).

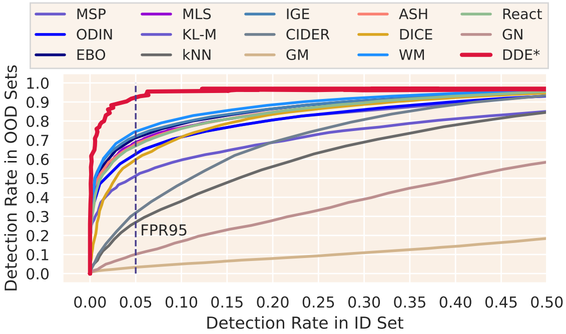


Figure 5: OOD detection rate is analyzed between ID set vs OOD sets, for WideResnet101. ID set is the figure is Imagenet test set and for OOD set, we take average across all the 51 OOD test sets. Detection rate in OOD sets should be high at the minimal cost of falsely detecting samples in the ID set as OOD. This plot demonstrates the superiority of *DDE** (our approach) w.r.t. all the competitive methods, it achieves OOD detection rate of 0.6 while not falsely detecting any of the ID samples as OOD, and the detection rates in OOD sets increases sharply for a very small increase in the false detection of OODs within the ID set.

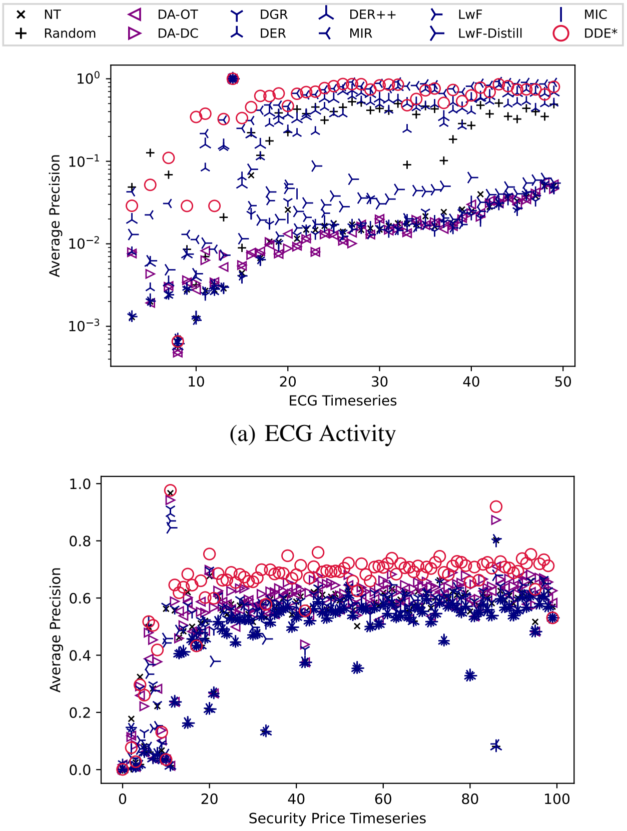
Methods	Summary Statistics from All Test Sets
ReAct	18±18
ASH	20±21
WM	20±22
DDE*	4±5
DDE-Online	5±6
DDE-Mixed	10±12
DDEv	12±17
DDEvt10	9±11
DDEvt20	7±8
DDE-N30k	9±9
DDE-N10k	14±11
DDE-N1k	15 ± 10
DDE-N100	31±14

Table 2: Evaluation of the variants of DDE* for OOD detection in WideResnet101 pretrained on Imagenet-1k using the metric FPR95 (↓).

In Fig. 4, we present the analysis from an extensive ablation study for our approach. First, in Fig. 4(a), we analyze how OOD detection rate increases in the ID set vs OOD test sets as we increase the number of batch updates of the weights of a neural KL-D estimator. As we discussed previously, a few hundred batch updates suffice in practice (considering a large batch size of 10k) for convergence of the KL-D estimates whereas performing a large number of batch updates (in thousands) can force a neural estimator to start distinguishing even between similar samples leading to divergence of the KL-D estimates. Correspondingly, in Fig. 4(a), we see that OOD detection rate within the ID set remains close to zero for the first few hundred iterations of batch updates. It is only if we keep on increasing the number of batch updates that the estimator starts detecting OOD samples even within the ID set, with 5% OOD detection rate in the ID set corresponding to the metric FPR95. In contrast, OOD detection rate across all the test OOD sets increases at a faster pace as it should be.

In 4(b), we analyze mean OOD detection across the test sets w.r.t. the number of batch updates, while varying the batch size. We find that using small batch size requires larger number of batch updates. The curves from batch sizes, 3k, 5k, 10k, 30k, are all alike in contrast to lower batch sizes. From theoretical standpoint, larger batch size is advantageous for achieving lower variance in the estimation of KL-D (note the zigzag in the curves for lower batch sizes). Note, the number of batch updates corresponding from FPR95 is different across the batch sizes and we only show FPR95 for the default batch size of 10k. In 4(c), we vary the number of hidden units. For all the three hidden sizes, 512, 1024, 2048, our approach performs similarly. The results for varying the learning rate are not surprising. Overall, it suggest that the effect of a hyperparameter on detection performance is intuitive and smooth.

In Table 2, we demonstrate competitiveness of the variants of DDE* w.r.t. best of the baselines (ReAct, ASH, WM). "DDE-Online" refers to batch inference on a test set. "DDE-Mixed" is for the evaluation setting of augmenting each OOD test set with (3000) ID test samples. For analyzing



(a) ECG Activity

(b) US Security Prices Activity

Figure 6: On the x-axis in each plot, we index individual timeseries within a dataset. For each timeseries, all the methods are compared in terms of Average Precision metric (↑).

generalization of the estimator, we optimize the dual function for estimating KL-divergence between the ID training set and the OOD validation set. Using this dual function, we perform OOD detection across all the OOD test sets. This highly compute efficient variant of our method is referred as "DDEv". Optionally, we fine tune for a given test set using 10% or 20% of the original compute cost of our method ("DDEvt10" and "DDEvt20"). We perform a new ablation study for our method (DDE*) by varying the sample size (N) on the Imagenet (ID) dataset; see "DDE-N30k", $\cdots$, "DDE-N100". See the supplement for more details.

3.2 IN-DISTRIBUTION DETECTION

Datasets and Evaluation Settings For the problem of ID detection, we consider the task of timeseries forecasting using two datasets, ECG Activity and US Security Price Activity. In the Security Price dataset of 1000 most liquid securities, given each of the 1000 securities, we augment the training set from the same security with (ID) samples from the historical past of the same security and of the other 999 securities. (Results reported for only the first 100 least liquid securities.) Same applies to ECG dataset of 50 timeseries. We preprocess each timeseries to obtain % change in activity w.r.t. the previous timestep. The task is of forecasting if the absolute value of % change is beyond a certain threshold (mean absolute value) given the knowledge of % change in the previous 100 timesteps. Evaluation metric is Average Precision (AP). See the supplement for more details.

Competitive Methods Various (contextual) replay techniques from the literature of continual learning are relevant as baselines for our method *DDE** (see Sec. 1.2 for details): (i) no transfer of knowledge via data augmentation (*NT*) (ii) random selection (*Random*); (iii) Learning without Forgetting (*LwF* and *LwF-Distill*); (iv) Dark Experience Replay (*DER* and *DER++*); (v) Deep Generative Replay (*DGR*); (vi) Maximally Interfered Retrieval (*MIR*); (vii) Memorable Information Criterion (*MIC*). Besides, from the literature of domain adaptation on invariant representation learning, we compare to (viii) domain discrimination (*DA-DC*), and (ix) neural optimal transport (*DA-OT*).

Empirical Results In the plots in In Fig. 6, X-axis represent indices of the individual timeseries in a given dataset. For each timeseries, we compare all the methods in terms of AP (↑). For ECG activity dataset, due to high interference between samples from different timeseries and high temporal dynamics within a single timeseries, we observe a very significant contrast between the methods. While *DDE** (ours) provides consistently highest AP across almost all the timeseries in the dataset, some of the other methods such as *DER++*, *MIR* are also competitive. For the dataset of US security price activity as well, *DDE** obtains the highest AP across all the timeseries, though the difference of AP across the methods is not as drastic as observed for ECG activity dataset. Another interesting aspect is that DA-DC is efficient for price activity in contrast to the ECG dataset.

4 CONCLUSIONS

In this paper, we tackle the highly impactful problem of OOD detection in pretrained DNNs. Our approach of OOD detection via dual divergence estimation is novel, principled, and highly efficient in practice. It enjoys theoretical guaranties owing to its foundations in information theory. While the approach is generic, one can employ a lightweight deep neural net as a dual function approximator for divergence estimation. Our extensive exprimental evaluation shows that our approach is drastically superior to all the competitive methods. We also establish benchmarks for a large number of new OOD test datasets. Moreover, we show that OOD detection is theoretically similar to ID detection, an underexplored problem with applications to continual learning and domain adaptation. For this problem as well, we provide theoretical guaranties and show its competitiveness w.r.t. many baselines on datasets from healthcare and finance domain.

References

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in Neural Information Processing Systems*, 2019.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 2020.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of International Conference on Machine Learning*, 2018.

Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems*, 2020.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, 2020.

Kamil Deja, Paweł Wawrzyński, Daniel Marczak, Wojciech Masarczyk, and Tomasz Trzciński. Binplay: A binary latent autoencoder for generative replay continual learning. In *International Joint Conference on Neural Networks*, 2021.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. i. *Communications on Pure and Applied Mathematics*, 1975.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022.

Ke Fan, Yikai Wang, Qian Yu, Da Li, and Yanwei Fu. A simple test-time method for out-of-distribution detection. *arXiv e-prints*, pages arXiv–2207, 2022.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2021.

David Freedman and Persi Diaconis. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 1981.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016.

Eduardo Dadalto Camara Gomes, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Igeood: An information geometry approach to out-of-distribution detection. In *Proceedings of International Conference on Learning Representations*, 2022.

Joris Guérin, Kevin Delmas, Raul Sena Ferreira, and Jérémie Guiochet. Out-of-distribution detection is not all you need. *arXiv preprint arXiv:2211.16158*, 2022.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proceedings of International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of International Conference on Machine Learning*, 2022.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 2021.

Wenjian Huang, Hao Wang, Jiahao Xia, Chengyan Wang, and Jianguo Zhang. Density-driven regularization for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2022.

Wenyu Jiang, Hao Cheng, Mingcai Chen, Shuai Feng, Yuxin Ge, and Chongjun Wang. Read: Aggregating reconstruction error into out-of-distribution detection. *arXiv preprint arXiv:2206.07459*, 2022.

Timothée Lesort, Massimo Caccia, and Irina Rish. Understanding continual learning settings with data distribution drift analysis. *arXiv preprint arXiv:2104.01678*, 2021.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of International Conference on Learning Representations*, 2018.

Jingyang Lin, Yu Wang, Qi Cai, Yingwei Pan, Ting Yao, Hongyang Chao, and Tao Mei. Out-of-distribution detection with hilbert-schmidt independence optimization. *arXiv preprint arXiv:2209.12807*, 2022.

Luping Liu, Yi Ren, Xize Cheng, and Zhou Zhao. Diffusion denoising process for perceptron bias in out-of-distribution detection. *arXiv preprint arXiv:2211.11255*, 2022.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing systems*, 2017.

Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. In *International Conference on Learning Representations*, 2020.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*. Elsevier, 1989.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *arXiv preprint arXiv:2211.13445*, 2022a.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *Proceedings of International Conference on Machine Learning*, 2022b.

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022c.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2021.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in Neural Information Processing Systems*, 2019.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2019.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *Proceedings of International Conference on Learning Representations*, 2018.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 2017.

Shengyang Sun, Daniele Calandriello, Huiyi Hu, Ang Li, and Michalis Titsias. Information-theoretic online memory selection for continual learning. In *Proceedings of International Conference on Learning Representations*, 2021.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Proceedings of European Conference on Computer Vision*, 2022.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of International Conference on Machine Learning*, 2022.

Binh Tang and David S Matteson. Graph-based continual learning. In *Proceedings of International Conference on Learning Representations*, 2020.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in Neural Information Processing Systems*, 2020.

Gido M Van de Ven and Andreas S Tolias. Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:1809.10635*, 2018.

Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 2020.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Qizhou Wang, Feng Liu, Yonggang Zhang, Jing Zhang, Chen Gong, Tongliang Liu, and Bo Han. Watermarking for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2022b.

Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, HAO Jianye, and Bo Han. Out-of-distribution detection with implicit outlier transformation. In *Proceedings of International Conference on Learning Representations*, 2023.

Yu Wang, Jingjing Zou, Jingyang Lin, Qing Ling, Yingwei Pan, Ting Yao, and Tao Mei. Out-of-distribution detection via conditional kernel independence model. In *Advances in Neural Information Processing Systems*, 2022c.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of International Conference on Machine Learning*, 2022.

Samuel Wilson, Tobias Fischer, Niko Sünderhauf, and Feras Dayoub. Hyperdimensional feature fusion for out-of-distribution detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Boxi Wu, Jie Jiang, Haidong Ren, Zifan Du, Wenxiao Wang, Zhifeng Li, Deng Cai, Xiaofei He, Binbin Lin, and Wei Liu. Towards in-distribution compatibility in out-of-distribution detection. *arXiv e-prints*, 2022.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022.

Mingtian Zhang, Andi Zhang, Tim Z Xiao, Yitong Sun, and Steven McDonagh. Ood detection with class ratio estimation. In *NeurIPS ML Safety Workshop*, 2022.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*, 2021.