

# Analyzing the Robustness of Open-World Machine Learning

Vikash Sehwa<sup>\*</sup>  
Princeton University

Arjun Nitin Bhagoji<sup>\*</sup>  
Princeton University

Liwei Song<sup>\*</sup>  
Princeton University

Chawin Sitawarin  
University of California, Berkeley

Daniel Cullina  
Pennsylvania State University

Mung Chiang  
Purdue University

Prateek Mittal  
Princeton University

## ABSTRACT

When deploying machine learning models in real-world applications, an *open-world* learning framework is needed to deal with both normal in-distribution inputs and undesired out-of-distribution (OOD) inputs. Open-world learning frameworks include OOD detectors that aim to discard input examples which are not from the same distribution as the training data of machine learning classifiers. However, our understanding of current OOD detectors is limited to the setting of benign OOD data, and an open question is whether they are robust in the presence of adversaries. In this paper, we present the first analysis of the robustness of open-world learning frameworks in the presence of adversaries by introducing and designing OOD adversarial examples. Our experimental results show that current OOD detectors can be easily evaded by slightly perturbing benign OOD inputs, revealing a severe limitation of current open-world learning frameworks. Furthermore, we find that OOD adversarial examples also pose a strong threat to adversarial training based defense methods in spite of their effectiveness against in-distribution adversarial attacks. To counteract these threats and ensure the trustworthy detection of OOD inputs, we outline a preliminary design for a robust open-world machine learning framework.

## CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Neural networks*; • **Security and privacy** → Intrusion/anomaly detection and malware mitigation;

## KEYWORDS

Open world recognition; Adversarial example, Deep learning

### ACM Reference Format:

Vikash Sehwa, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. 2019. Analyzing the Robustness of Open-World Machine Learning. In *12th ACM Workshop on Artificial Intelligence and Security (AISec '19)*, November 15, 2019, London, UK. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3338501.335737>

<sup>\*</sup>Equal contribution

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*AISec '19*, November 15, 2019, London, United Kingdom

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6833-9/19/11.

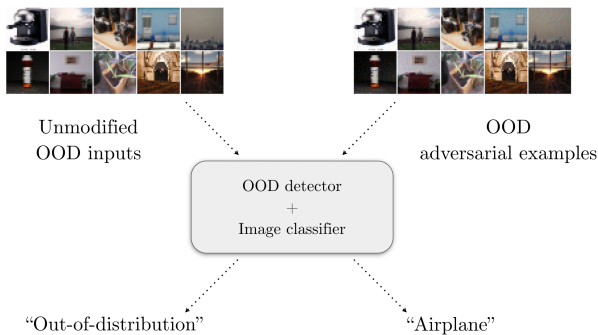
<https://doi.org/10.1145/3338501.335737>

## 1 INTRODUCTION

Machine learning (ML) models, especially deep neural networks, have become prevalent and are being widely deployed in real-world applications, such as image classification [40, 57], face recognition [51, 61], and autonomous driving [9, 16]. Motivated by the fact that real-world applications need to be resilient to arbitrary input data, an important line of work has developed the open-world learning framework that checks if the inputs are within the same distribution as training data (in-distribution examples), or if they come from a different distribution referred to as *out-of-distribution* (OOD) examples [5, 6]. State-of-the-art open-world learning systems equip machine learning classifiers with OOD detectors, and an input example is processed for classification only if the input passes through those detectors. In recent years, the research community has developed several OOD detection mechanisms that are effective in distinguishing OOD inputs [30, 42, 44].

However, a severe limitation of current open-world learning frameworks is that their development and investigation has been limited to the setting of benign (natural and unmodified) OOD data. Despite their good performance in detecting benign OOD inputs, an important open question is *whether open-world learning frameworks are robust in the presence of adversaries?* Specifically, can OOD detectors perform reliably when an adversary tries to evade them by maliciously perturbing OOD inputs? In this paper, we thoroughly evaluate the performance of open-world machine learning models against such maliciously-perturbed OOD inputs, which we refer to as *OOD adversarial examples*, motivated by the line of research on adversarial attacks against neural networks [8, 11, 62]. Our analysis shows that state-of-the-art OOD detectors [42, 44] are quite fragile: their detection performance drops drastically with perturbations to out-of-distribution inputs. For example, as highlighted in Figure 1, benign OOD inputs can be reliably detected as out-of-distribution by current open-world learning systems. However, OOD adversarial examples are able to both evade the OOD detector as well as achieve targeted misclassification by the classifier.

Beyond revealing the lack of robustness of current OOD detectors, we further examine the behavior of OOD adversarial examples on state-of-the-art robustly trained classifiers [46, 65], which were designed for robustness against in-distribution adversarial examples. This *novel examination is critical because once the adversary manages to bypass the OOD detector, the open-world learning framework will pass that input to the relevant classifier*. We find that compared to in-distribution attacks, OOD adversarial examples result in much higher attack success rates against robust classifiers.



**Figure 1: Performance of an open-world image classification system (OOD detector + image classifier) on unmodified OOD inputs and OOD adversarial examples for Wide-ResNet-28-10 [69] classifier trained on CIFAR-10 along with ODIN [44] as OOD detector. While unmodified OOD inputs are easily detected by the detector, OOD adversarial examples can evade the detector and achieve targeted classification (e.g., “airplane”) with confidence close to 100%.**

Under the taxonomy of attacks on ML systems laid out by Huang et al. [34], OOD adversarial examples are a form of *exploratory integrity attacks* with the intent of *targeted* misclassification of input data. Intuitively, we would expect a well-behaved classifier to classify an OOD adversarial example with low confidence, since it has never encountered an example from that portion of the input space before. This intuition underlies the design of state-of-the-art OOD detectors. However, OOD adversarial examples constitute an integrity violation for classifiers as well as OOD detectors as they induce high confidence targeted misclassification in state-of-the-art classifiers, which is unwanted system behavior.

While previous work [26, 50, 56, 58] has hinted at the possibility of generating adversarial examples without the use of in-distribution data, we are the first to rigorously examine the impact of OOD adversarial examples on OOD detectors and robustly trained ML classifiers. We also showcase their feasibility in real-world contexts by demonstrating attacks against the Clarifai content moderation system [17] and a traffic sign classification system. Finally, to counteract these threats, we also outline a preliminary design for robust open-world learning by combining OOD adversarial examples with adversarial training, and demonstrate the generalization properties of this approach.

In summary, we make the following contributions in this paper: **Robustness evaluation of open-world learning framework:** We introduce and design OOD adversarial examples by adversarially perturbing OOD inputs for evading OOD detectors. Our experiments with two state-of-the-art OOD detection mechanisms, ODIN [44] and Confidence-calibrated classifiers [42], show that current open-world machine learning models are not robust: most OOD inputs (up to 99.8%) can pass through OOD detectors successfully by adding imperceptible perturbations.

**Bypassing state-of-the-art defenses for in-distribution attacks:** Although state-of-the-art defenses such as iterative adversarial training [46] and provably robust training with the convex outer polytope [38, 65] are promising approaches for the mitigation of in-distribution attacks, their performance significantly degrades

with the use of OOD adversarial examples. We demonstrate that OOD adversarial examples can achieve a significantly higher target success rate (up to 4× greater) than that of adversarial examples generated from in-distribution data. Further, we demonstrate that OOD adversarial examples are able to evade adversarial example detectors such as feature squeezing [68] and MagNet [48], with close to 100% success rate (similar to in-distribution adversarial examples). We also show this for Adversarial Logit Pairing [37].

**OOD adversarial examples in the real world:** We demonstrate the success of OOD adversarial examples in real-world settings by targeting a content moderation service provided by Clarifai [17]. We also show how physical OOD adversarial examples can be used to fool traffic sign classification systems.

**Towards robust open-world learning:** We explore the possibility of increasing the robustness of open-world machine learning by including a small number of OOD adversarial examples in robust training. Our results show that such an increase in robustness, even against OOD datasets excluded in training, is possible.

We hope that our work serves to inspire a rigorous understanding of open-world learning frameworks in the presence of adversaries, with the end-goal of facilitating trustworthy and safe deployment of open-world ML systems<sup>1</sup>.

## 2 BACKGROUND AND RELATED WORK

In this section we present the background and related work on open-world deep learning, adversarial examples generated from in-distribution data, and corresponding defenses.

### 2.1 Supervised classification

Let  $\mathcal{X}$  be a space of examples and let  $\mathcal{Y}$  be a finite set of classes. A classifier is a function  $f(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ . Let  $\mathbb{P}(\mathcal{Y})$  be the set of probability distributions over  $\mathcal{Y}$ . In our setting, a classifier is always derived from a function  $g(\cdot) : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{Y})$  that provides confidence information, i.e.  $f(x) = \operatorname{argmax}_{i \in \mathcal{Y}} g(x)(i)$ . In particular, for DNNs, the outputs of the penultimate layer of a neural network  $f$ , representing the output of the network computed sequentially over all preceding layers, are known as the logits. We represent the logits as a vector  $\phi^f(x) \in \mathbb{R}^{|\mathcal{Y}|}$ . The classifier is trained by minimizing the empirical loss  $\frac{1}{n} \sum_{i=1}^n \ell_g(x_i, y_i)$  over  $n$  samples  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  (training set), where  $\ell_g(\cdot, \cdot)$  is a loss function such as the cross-entropy loss [25] that depends on the output confidence function  $g(\cdot)$ . The training set is drawn from a distribution  $P_{\mathcal{X}, \mathcal{Y}}^{\text{in}}$  over the domain  $\mathcal{X} \times \mathcal{Y}$ . The marginal distribution over the space of examples  $\mathcal{X}$  is represented as  $P_{\mathcal{X}}^{\text{in}}$ . These samples usually represent an application-specific set of concepts that the classifier is being trained for.

### 2.2 Open-world Deep learning

The closed-world approach to deep learning, described in Section 2.1 operates using the assumption that both training and test data are drawn from the same application-specific distribution  $P_{\mathcal{X}, \mathcal{Y}}^{\text{in}}$ . However, in a real-world environment, ML systems need to be resilient to data at test time that is not drawn from  $P_{\mathcal{X}, \mathcal{Y}}^{\text{in}}$  but belongs to the same input space, i.e. they encounter samples that are *out-of-distribution* (OOD). This leads to the *open-world learning model*.

<sup>1</sup>An extended technical report [55] with additional results and our code (<https://github.com/inspire-group/OOD-Attacks>) are available.

Thus, in order to extend supervised learning algorithms to the open-world learning model, it is critical to enable them to reject out-of-distribution inputs. The importance of this learning model is highlighted by the fact that a number of security and safety-critical applications such as biometric authentication, intrusion detection, autonomous driving, medical diagnosis are natural settings for the use of open-world machine learning [13, 27, 45, 53].

**2.2.1 Out-of-Distribution data.** To design and evaluate the success of an open-world ML approach, it is first critical to define out-of-distribution data. Existing work on open-world machine learning [5, 6, 20] defines an example  $\mathbf{x}$  as OOD if it is drawn from a marginal distribution  $P_X^{\text{out}}$  (over  $\mathcal{X}$ , the input feature space) which is different from  $P_X^{\text{in}}$  and has a label set that is disjoint from that of in-distribution data. As a concrete example, consider a classifier trained on the CIFAR-10 image dataset [39]. This dataset only has 10 output classes and does not include classes for digits such as ‘3’ or ‘7’ like the MNIST [41] dataset or ‘mushroom’ or ‘building’ like the Imagenet dataset [19]. Thus, these datasets can act as a source of OOD data.

**2.2.2 OOD detectors.** Here, we only review recent approaches to OOD detection that scale to DNNs used for image classification. Hendrycks and Gimpel [30] proposed a method for detecting OOD inputs for neural networks which uses a threshold for the output confidence vector to classify an input as in/out-distribution. This method relies on the assumption that the classifier will tend to have higher confidence values for in-distribution examples than OOD examples. An input is classified as being OOD if its output confidence value is smaller than a certain learned threshold.

In this work, we evaluate the state-of-the-art OOD detectors for DNNs proposed by Liang et al. [44] (ODIN) and Lee et al. [42] which also use output thresholding for OOD detection but significantly improve upon the baseline approach of Hendrycks and Gimpel [30]. The ODIN detector uses temperature scaling and input pre-processing to improve detection rates. Lee et al. [42] propose a modification to the training procedure to ensure that the neural network outputs a confidence vector which has probabilities uniformly distributed over classes for OOD inputs. However, as OOD inputs are unavailable at training time, they generate synthetic data, using a modified Generative Adversarial Network (GAN), which lies on the boundary between classes to function as OOD data.

## 2.3 Evasion attacks and defenses

**2.3.1 Evasion attacks.** Evasion attacks are test-time attacks that have been demonstrated to be highly successful for a number of ML classifiers [8, 11, 26, 62]. These attacks aim to modify benign, *in-distribution* examples  $\mathbf{x} \sim P_X^{\text{in}}$  by adding an imperceptible perturbation to them such that the modified examples  $\tilde{\mathbf{x}}$  are *adversarial* [62]. The adversary’s aim is to ensure that these adversarial examples are successfully misclassified by the ML system in a targeted class (targeted attack), or any class other than the ground truth class (untargeted attack). We focus entirely on *targeted attacks* since these are more realistic from an attacker’s perspective and are strictly harder to carry out than untargeted attacks. To generate a successful *targeted* adversarial example  $\tilde{\mathbf{x}}$  for class  $T$  starting from a benign example  $\mathbf{x}$  for a classifier  $f$ , the following optimization

problem must be solved

$$f(\tilde{\mathbf{x}}) = T \quad \text{s.t.} \quad d(\tilde{\mathbf{x}}, \mathbf{x}) < \epsilon \quad (1)$$

where  $d(\cdot, \cdot)$  is an appropriate distance metric for inputs from the input domain  $\mathcal{X}$  used to model imperceptibility-based adversarial constraints [11, 26]. The distance metric imposes an  $\epsilon$ -ball constraint on the perturbation. The optimization problem in Eq. 1 is combinatorial and thus difficult to solve. In practice, a relaxed version using an appropriate adversarial loss function  $\ell_g^{\text{adv}}(\mathbf{x}, T)$  derived from the confidence function  $g(\cdot)$  is used and solved with an iterative optimization technique [11, 26, 46, 62]. Details of the state-of-the-art attack methods we use are in Section 4.3.

The two key threat models for these adversarial attacks are white-box and black-box with query access. While in the former the adversary has complete access to the classifier, including employed defenses, dataset, and hyperparameters, the latter threat model only allows access to the output probability distribution  $g(\cdot)$  for any input  $\mathbf{x} \in \mathcal{X}$  [3, 11, 26, 62]. Work on adversarial examples has also examined the threat they pose in real-world settings. One line of work has been to analyze attacks on real-world ML services in black-box threat models [7, 35] while other focus on breaking ML systems with physically realized adversarial examples [4, 23]. We will engage with both of these directions in Section 5.4.

Current research on evasion attacks is limited in the closed-world setting: adversarial examples are generated using training or test inputs as starting points. Meanwhile, as shown in our paper, extending evasion attacks to the out-of-distribution space is natural and critical for the robustness evaluation of open-world machine learning systems.

**2.3.2 Defenses against evasion attacks.** Robust training is one of the most effective methods to achieve robustness against adversarial attacks on deep neural networks [26, 38, 46, 52, 64]. It seeks to *embed resilience* into the classifier during training by modifying the standard loss  $\ell_g(\cdot, \cdot)$  used during the training process to one that accounts for the presence of an adversary at test time. It can be divided into two categories:

**Adversarial training:** These heuristic methods defend against adversarial examples by modifying the loss function such that it incorporates both clean and adversarial inputs [26, 37, 46].

$$\tilde{\ell}_g(\mathbf{x}, y) = \alpha \ell_g(\mathbf{x}, y) + (1 - \alpha) \ell_g(\tilde{\mathbf{x}}, y), \quad (2)$$

where  $y$  is the true label of the sample  $\mathbf{x}$ .

In this paper, we consider the robustness of networks trained using *iterative adversarial training* [46, 67], which uses adversarial examples generated using Projected Gradient Descent (PGD). This method has been shown to be empirically robust to adaptive white-box adversaries using adversarial examples generated from in-distribution data which use the same  $L_p$  norm [3] for models trained on the MNIST [41], CIFAR-10 [39], and ImageNet [19] datasets.

**Provable robustness using convex relaxations:** We focus on the approach of Kolter and Wong [38, 65] which scales to neural networks on the CIFAR-10 dataset [39]. They aim to certify robustness in an  $\epsilon$ -ball around any point in the input space by upper bounding the adversarial loss with a convex surrogate. They find a convex outer approximation of the activations in a neural network that can be reached with a perturbation bounded within an  $\epsilon$ -ball

and show that an efficient linear program can be used to minimize the worst case loss over this region.

In Section 5.2, we show reduced effectiveness of both adversarial training and provable robustness based defenses for OOD adversarial examples.

**2.3.3 Adversarial Example Detectors and secondary defenses.** Adversarial detectors [48, 54, 68] aim to exploit the difference in properties of adversarial and unmodified input to detect adversarial examples. We consider two of the most promising detectors, namely *feature squeezing* [68] and *MagNet* [48]. In Section 5.3, we show that these detectors lack robustness to OOD adversarial examples.

Initial works on iterative adversarial training [46, 64] have highlighted the challenge of scaling and convergence for Imagenet-scale datasets [46]. The first approach to overcome this challenge was *Adversarial Logit Pairing* (ALP) [37]. However, it was shown that simply increasing the number of PGD attack iterations used to generate adversarial examples from in-distribution data reduced the additional robustness to a negligible amount [21]. In Section 5.2, we show that this lack of robustness persists for OOD adversarial examples.

### 3 OPEN-WORLD EVASION ATTACKS

Deployed ML systems must be able to robustly handle inputs which are drawn from distributions other than those used for training/testing (open-world learning [42, 44]). In order to test the robustness of such open-world learning systems, we define *open-world evasion attacks*, which can use arbitrary points from the input space to generate *out-of-distribution (OOD) adversarial examples*, making our work the first to combine the paradigms of open-world learning and adversarial examples. We then analyze the effectiveness of OOD adversarial examples in bypassing OOD detectors as well as defenses for in-distribution adversarial examples. We find that OOD adversarial examples present a potent threat in the open-world learning model, and summarize our key results in Table 1. Finally, we also examine how robustness against OOD adversarial examples can be achieved.

#### 3.1 OOD adversarial examples

In the open-world learning model, an adversary is not restricted to in-distribution data, and can generate adversarial examples using OOD data. In order to carry out an evasion attack in this setting, the adversary generates an *OOD adversarial example* starting from  $\mathbf{x}_{\text{OOD}}$ .

**Definition 1** (OOD adversarial examples). An OOD adversarial example  $\tilde{\mathbf{x}}_{\text{OOD}}$  is generated from an OOD example  $\mathbf{x}_{\text{OOD}}$  drawn from  $P_X^{\text{out}}$  by adding a perturbation  $\delta$  with the aim of inducing classification in a target class  $T \in \mathcal{Y}_1$ , i.e.  $f(\tilde{\mathbf{x}}_{\text{OOD}}) = T$ .

Existing attack methods use optimization-based approaches [3, 11, 26, 62] to generate in-distribution adversarial examples, which are misclassified with high confidences. While in the open-world learning setting, we also need to consider OOD detection mechanisms in the attack algorithm for detection bypassing (see further discussion in Section 3.1.1). Next, we highlight the importance of

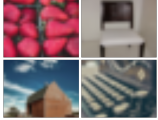
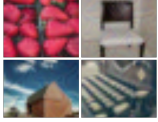
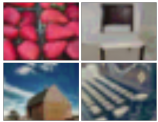
constructing OOD adversarial examples to fool classifiers by discussing the limitations of directly using unmodified/benign OOD data.

**Limitations of unmodified OOD data for evasion attacks:** While unmodified OOD data already represents a concern for the deployment of ML classifiers (Section 2.2), we now discuss why they are severely limited from an adversarial perspective. First, with unmodified OOD data, the typical output confidence values are small, while an attacker aims for high-confidence targeted misclassification. Second, the attacker will have no control over the target class reached by the unmodified OOD example. Finally, due to the low typical output confidence values of unmodified OOD examples they can easily be detected by state-of-the-art OOD detectors (Section 2.2.2) which rely on confidence thresholding.

**3.1.1 Evading OOD detectors.** OOD detectors are an essential component of any open-world learning system and OOD adversarial examples should be able to bypass them in order to be successful. State-of-the-art OOD detectors mark inputs which have confidence values below a certain threshold as being OOD. The intuition is that a classifier is more likely to be highly confident on in-distribution data. Recall that when generating adversarial examples, the adversary aims to ensure high-confidence misclassification in the desired target class. The goal of an adversary seeking to generate high-confidence targeted OOD adversarial examples will align with that of an adversary aiming to bypass an OOD detector. In other words, *OOD adversarial examples that achieve high-confidence targeted misclassification also bypass OOD detectors*. Our empirical results in Section 5.1 demonstrate this conclusively, with OOD adversarial examples inducing high false negative rates in the OOD detectors, which mark them as being in-distribution.

**3.1.2 Evading robust training based defenses.** Given the lack of robustness of OOD detectors, it becomes important to examine the impact of OOD adversarial examples on the underlying (in-distribution) classifier. This is because in an open-world learning framework, input examples which pass through OOD detectors are processed by the underlying classifier for a final prediction. For our analysis, we focus on robustly trained neural networks as the underlying classifier. Robustly trained neural networks [38, 46, 52] (recall Section 2.3.2), incorporate the evasion attack strategy into the training process. Since the training and test data are drawn in an i.i.d. fashion, the resulting neural networks are robust to in-distribution adversarial examples at test time. However, these networks may not be able to provide robustness if the attack strategy were to be changed. In particular, we change the starting point for the generation of adversarial examples to be out-of-distribution samples, and since *the training process for these robust defenses does not account for the possibility of OOD data being encountered at test time, they remain vulnerable to OOD adversarial examples*. We find that for defenses based on robust training, OOD adversarial examples are able to increase targeted success rate by 4× (Sections 5.2.2 and 5.2.3). This finding illustrates the potent threat of open-world evasion attacks, which must be addressed for secure deployment of ML models in practice. We further demonstrate that adversarial example detectors such as MagNet and Feature Squeezing can be similarly bypassed by incorporating the metrics and pre-processing they use into the attack objective for OOD adversarial examples.

**Table 1: Summary of results on CIFAR-10 trained models. Novel results and empirical conclusions from this paper are in bold. The last column shows the successful OOD adversarial examples for each defense, where all images are classified as *airplane*.**

Defense Type	Defense Name	Behavior on data type		Representative OOD adversarial examples
		In-distribution adversarial (white-box, adaptive)	<b>Out-of-distribution adversarial</b> <b>(white-box, adaptive)</b>	
N.A.	Undefended	Not robust [11, 46] Rate: 100.0, Conf: 1.00	Not robust Rate: 100.0, Conf: 1.00 (ImageNet)	
OOD Detection	ODIN [44]	N.A.	Not robust Rate: 81.6, Conf: 0.97 (Internet Photographs)	
	Confidence-calibrated [42]	N.A.	Somewhat robust Rate: 47.1, Conf: 0.99 (VOC12)	
Robust Training	Iterative adv. training [46]	Somewhat robust [46] Rate: 22.9, Conf: 0.81	Not robust Rate: 87.9, Conf: 0.86 (Gaussian Noise)	
	Convex polytope relaxation [38]	Provably robust [38] Rate: 15.1, Conf: 0.41	Somewhat robust Rate: 29.1, Conf: 0.32 (Gaussian Noise)	

**3.1.3 Real-world attacks.** Since the aim of the open-world threat model is to elucidate the wider range of possible threats to a deployed ML model than previously considered, we analyze the possibility of realizing OOD adversarial examples in the following real-world settings:

- (1) **Physical attacks:** We consider attacks on a traffic sign recognition system where an adversary uses custom signs and logos in the environment as a source of OOD adversarial examples, since the classifier has only been trained on traffic signs. In a physical setting, there is the additional challenge of ensuring that the OOD adversarial examples remain adversarial in spite of environmental factors such as lighting and angle. We ensure this by incorporating random lighting, angle and re-sizing transformations into the OOD adversarial example generation process [4, 23].
- (2) **Query-limited black-box attacks:** We use OOD adversarial examples to carry out a Denial of Service style attack on a content moderation model provided by Clarifai [17], by *classifying clearly unobjectionable content as objectionable with high confidence*. Since we only have query-access to the model being attacked, the model gradients usually needed to generate adversarial examples (see Section 4.3) have to be estimated. This is done using the finite difference method with random grouping based query-reduction [7].

Our results in Section 5.4 show that OOD adversarial examples remain effective in these settings and are able to successfully attack content moderation and traffic sign recognition systems.

## 3.2 Towards Robust Open-World Deep Learning

A robust open-world deep learning system is expected to satisfy the following two properties: (i) It should have high accuracy in detecting both unmodified and adversarial OOD inputs; (ii) It should

have high accuracy in classifying unmodified and adversarial in-distribution inputs. To move towards a robust open-world deep learning system, we take inspiration from previous work on selective prediction [15, 24, 47] which augments classifiers for in-distribution data with an additional class (referred to as background class) so they can be extended to open-world learning environment and detect OOD inputs. Further, since iterative adversarial training [46] enables the robust classification of in-distribution adversarial examples, we can intuit that a similar approach may provide robustness to OOD adversarial examples. Thus, we examine a *hybrid approach where we use iterative adversarial training to ensure robust classification of OOD data, both unmodified and adversarial, to the background class*. Similar to other OOD detection approaches [30, 31, 42, 44], selective prediction is semi-supervised, i.e. it assumes access to a small subset of OOD data at training time. As highlighted by our results in Section 6, this detector can successfully generalize to unseen OOD adversarial examples. We note that since all of these state-of-the-art approaches consider the detection of specific (multiple) OOD datasets, we follow the same methodology for robust OOD classification. To achieve robust classification in the open-world environment, we perform iterative adversarial training with the following loss function:

$$\tilde{\ell}_g(\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{OOD}}, y) = \alpha \tilde{\ell}_g(\mathbf{x}_{\text{in}}, y) + (1 - \alpha) \tilde{\ell}_g(\mathbf{x}_{\text{OOD}}, y_b) \quad (3)$$

where  $\mathbf{x}_{\text{in}} \in P_X^{\text{in}}$ ,  $\mathbf{x}_{\text{OOD}} \in P_X^{\text{out}}$ ,  $y$  is true label for sample  $\mathbf{x}_{\text{in}}$  and  $y_b$  is the background class.  $\tilde{\ell}_g(\cdot, \cdot)$  refers to the robust loss used in adversarial training (Eq. 2).

The question now arises: to what extent does this formulation satisfy the two desired properties from a robust open-world learning system? In particular, we examine if the following goals are feasible using small subsets of OOD data: i) robust classification of a single OOD dataset? ii) generalization of robustness to multiple OOD datasets while training with a single one? iii) simultaneous

robustness to multiple OOD datasets while training with data from all of them? Again, we emphasize that these must be achieved while maintaining high accuracy on in-distribution data.

Our evaluation in Section 6 answers these questions in the affirmative. For example, we observe that a subset as small as 0.5% of the total number of samples from an OOD dataset can significantly enhance robustness against OOD adversarial examples.

## 4 DESIGN CHOICES FOR OPEN-WORLD EVASION ATTACKS

In this section, we present and examine the design choices we make to carry out our experiments on both evaluation and training of classifiers in the open-world model. In particular, we discuss the types of datasets, models, attack methods, and metrics we consider. All our code and links to the data is available on <https://github.com/inspire-group/OOD-Attacks> for the purposes of reproducible research.

### 4.1 Datasets

We consider 3 publicly available datasets for image classification as sources of in-distribution data for training ( $P_{X,Y}^{\text{in}}$ ): MNIST [41], CIFAR-10 [39], and ImageNet [19]. When one of the above datasets is used for training, the other two datasets are used as sources of OOD data. We consider the following two types of OOD data.

**Semantically meaningful OOD data:** Datasets such as MNIST, CIFAR-10 and ImageNet are *semantically meaningful* as the images they contain generally have concepts recognizable to humans. To further explore the space of semantically meaningful OOD data, we also consider the VOC12 [22] dataset as a source of OOD data. Furthermore, we construct an Internet Photographs dataset by gathering 10,000 natural images from the internet using the Picsum service [2]. To avoid any ambiguity over examples from different datasets that contain similar semantic information, we ensure the label set for semantically meaningful OOD examples is distinct from that of the in-distribution dataset.

**Noise OOD data:** By definition, OOD data does not have to contain recognizable concepts. Thus, we construct a Gaussian Noise dataset consisting of 10,000 images for each of which the pixel values are sampled from a Gaussian distribution with the mean value equal to 127, and the standard deviation equal to 50. In settings where inputs to an ML classifier are not checked for any sort of semantics, this dataset is a viable input and thus must be analyzed when checking for robustness.

### 4.2 Models

We experiment with three robust training defenses (Iterative adversarial training [46], Adversarial logit pairing [37], and robust training with convex outer polytope [38]), two adversarial example detectors (Feature Squeezing [68] and MagNet [48]), and two OOD detectors (ODIN [44] and Confidence calibrated classifiers [42]). The details about model architectures and classification performance are described in Table 2. Following the convention in previous work [11, 38, 46], we report the perturbation budget for models trained on MNIST dataset using [0,1] scale instead of [0,255].

**Table 2: Deep neural networks used for each dataset in this work. The top-1 accuracy and confidence values are calculated using the test set of the respective datasets.**

Dataset	Model	Classification accuracy (%)	Mean confidence
MNIST [41]	4-layer CNN( $M_1$ ) [46]	98.8	0.98
	4-layer CNN ( $M_2$ ) [38]	98.2	0.98
	7-layer CNN ( $M_3$ ) [11]	99.4	0.99
CIFAR-10 [39]	Wide Residual net (WRN-28-10) [69]	95.1	0.98
	WRN-28-10-A [46, 69]	87.2	0.93
	WRN-28-1 [65]	66.2	0.57
	DenseNet [33]	95.2	0.98
	VGG13 [57]	80.1	0.94
	All Convolution Net [59]	85.7	0.96
ImageNet [19]	MobileNet [32]	70.4	0.71
	ResNet-v2-50 [28, 37]	60.5	0.28
	(for 64×64 size images)		

### 4.3 OOD Evasion Attack Methods

**4.3.1 Targeted adversarial attacks.** We only consider *targeted* attacks for two reasons. First, they are strictly harder than non-targeted attacks for the adversary [11]. Second, unmodified OOD examples have no ground truth labels, which raises difficulties in defining non-targeted attacks and comparing them to the in-distribution case. We select the target label randomly for each input image from the set of possible output classes *excluding the current predicted class*.

**4.3.2 Distance constraints.** In this paper, we use the  $L_\infty$  perturbation constraint for most of our attacks, except for the attack on feature squeezing [68], where the  $L_\infty$  metric cannot be used due to bit depth reduction and thus the  $L_2$  perturbation is adopted instead. These metrics are widely used to generate in-distribution adversarial examples because examples  $x$  and  $\tilde{x}$  that are  $\epsilon$ -close in these distance metrics, can have similar visual semantics [11, 26, 62].

#### Why use distance constraints for OOD adversarial examples?

There are two main reasons why we use distance constraints to generate OOD adversarial examples. The first reason, which *applies only to semantically meaningful OOD data*, is to model the content in the input that the adversary wishes to preserve, in spite of it being OOD. In other words, the starting point itself models a constraint on the adversary, which may arise from the environment (see Section 5.4.2 for an example for traffic sign recognition systems) or to prevent the OOD adversarial example from having undesirable artifacts, e.g. turning a non-objectionable image into an objectionable one (see Section 5.4.1).

The second reason, which applies to both semantically meaningful and noise OOD data, stems from the *need to measure the spread of successful OOD adversarial examples in the input space*. Previous work has measured the spread of adversarial examples around the training and test data, in terms of  $L_p$  distance constraints and found that for undefended models adversarial examples are present close to their starting points. While the use of robust training defenses makes it challenging to find adversarial examples within a given constraint set, we show for OOD data, successful OOD adversarial

examples can be found in small  $L_p$  balls around unmodified starting points for both undefended and defended models. We note that for noise OOD data it is possible to relax distance constraints to generate OOD adversarial examples which will lead to higher attack success rates. In Section 5.4.2, we demonstrate open-world evasion attacks using custom signs on traffic sign recognition systems that do not restrict the perturbation budget as well.

**4.3.3 Attack algorithm.** For an OOD input  $\mathbf{x}_{\text{OOD}}$  with a target label  $T$ , we choose *Projected Gradient Descent* (PGD) algorithm to generate OOD adversarial examples, since it presents state-of-the-art adversarial attack performance. PGD algorithm iteratively minimizes the target prediction loss  $\ell_g(\cdot, T)$  and then project onto the constraint set  $\mathcal{H}$  to follow the  $L_p$  perturbation constraint, which can be expressed as

$$\tilde{\mathbf{x}}_{\text{OOD}}^t = \Pi_{\mathcal{H}}(\tilde{\mathbf{x}}_{\text{OOD}}^{t-1} - \alpha \cdot \text{sign}(\nabla_{\tilde{\mathbf{x}}^{t-1}} \ell_g^{\text{adv}}(\tilde{\mathbf{x}}_{\text{OOD}}^{t-1}, T))), \quad (4)$$

where  $t$  is the total number of steps,  $\Pi$  is a projection operator,  $\tilde{\mathbf{x}}_{\text{OOD}}^0 = \mathbf{x}_{\text{OOD}}$  and  $\tilde{\mathbf{x}}_{\text{OOD}}^t$  is the final adversarial example. In our experiments,  $t$  is usually set to be 100-1000, and we choose an appropriate step size  $\alpha$  based on adversarial perturbation constraint to obtain the attack performance. For the target prediction loss  $\ell_g(\cdot, T)$ , we use the standard *cross-entropy* [25] loss function. As described in Section 3.1.1, the OOD detectors [42, 44] considered in our paper compare prediction confidence with a threshold to detect OOD inputs. Since the attack algorithm in Equation (4) iteratively increases prediction confidence by decreasing the targeted prediction loss, it automatically bypasses the OOD detectors.

*Note.* While most of our experiments are for adaptive white-box adversaries, in Section 5.4 we use query-based black-box attacks which rely on PGD but due to a lack of access to the true gradient, we use estimated gradients instead.

## 4.4 Evaluation Metrics

We consider the following metrics to measure the performance and robustness of OOD detectors and image classifiers:

**Classification accuracy:** This is the percentage of in-distribution test data for which the predicted label matches the ground-truth label. It is not reported for OOD data as they have no ground truth labels.

**False negative rate:** This is the percentage of OOD examples by-passing OOD detectors in the open-world learning framework, which measures the performance of detectors.

**Target success rate:** This is the percentage of adversarial examples classified as the desired target, which measures model robustness.

**Mean classification confidence:** This is the mean value of the output probability corresponding to the predicted label of the classifier on correctly classified inputs from in-distribution data. For adversarial examples, both in-distribution and OOD, it is the mean value of the output probability corresponding to the target label for successful adversarial examples. The confidence values lie in  $[0, 1]$ .

## 5 RESULTS

In this section, we present the experimental results. We first evaluate the robustness of OOD detectors to OOD adversarial examples. Next, we analyze the performance of state-of-the-art defenses against adversarial examples in the open-world learning model. We

**Table 3: False Negative rate (FNR) of ODIN [44] and confidence calibrated classifier [42] approaches for OOD adversarial examples. The results are reported with the respective models for each detector trained on CIFAR-10 dataset. The TNR of each detector with in-distribution dataset is 95%. These results show that a high percentage of OOD adversarial examples can evade OOD detectors.**

OOD dataset	ODIN [44]		Confidence-calibrated classifier [42]	
	$\epsilon = 8.0$	$\epsilon = 16.0$	$\epsilon = 8.0$	$\epsilon = 16.0$
ImageNet	68.8	97.4	46.4	<b>47.5</b>
VOC12	74.4	97.4	<b>47.1</b>	47.2
Internet	<b>81.6</b>	98.7	42.5	45.4
Photographs				
MNIST	72.6	<b>99.8</b>	4.6	5.2
Gaussian	0	4.2	20.9	21.9
Noise				

also evaluate adversarial detectors and secondary defenses in this setup. Finally, we demonstrate two open-world evasion attack cases with real-world attacks using OOD adversarial examples.

### 5.1 OOD detectors are not robust

In this section, we evaluate the robustness of two OOD detectors, ODIN [44] and confidence calibrated classification [42], to OOD adversarial examples with models trained on CIFAR-10 dataset.

**OOD detector setup.** The success of OOD adversarial examples against an OOD detector is measured by False Negative Rate (FNR), which represents the fraction of OOD inputs the detector fails to detect. The threshold values reported in [42, 44] are calibrated such that the True Negative Rate (TNR) i.e., the fraction of in-distribution inputs the detector classifies as non-OOD, is equal to 95%.

**5.1.1 Effect of OOD attacks on ODIN.** We use the code and the pre-trained DenseNet [33] model on CIFAR-10 dataset from Liang et al. [44]. For consistency, we follow Liang et al. [44] and use the temperature scaling and perturbation budget for input pre-processing equal to 1000 and 0.0014 respectively.

**OOD adversarial examples can evade the ODIN detector successfully.** We first test the performance of ODIN with multiple unmodified OOD datasets. As expected, ODIN achieves more than 78% detection accuracy for all unmodified OOD datasets. However, *the detection rate of ODIN drastically decreases with OOD adversarial examples* (Table 3). Except for Gaussian Noise, the mean attack success rate is 98.3% for other OOD datasets with  $\epsilon = 16$ .

**5.1.2 Effect of OOD attacks on confidence calibrated classifiers.** For consistency with prior work of Lee et al. [42], we use a similar model (VGG13) and training procedure. We also validate the results from Lee et al. [42] by evaluating it on unmodified OOD datasets. **Up to 47.5% OOD adversarial examples could bypass the detection approach based on confidence calibrated classifiers.** We first found that the confidence-calibrated classifier has good detection performance on unmodified OOD data. For example, unmodified ImageNet and VOC12 dataset are correctly detected with

accuracy higher than 80%. However, the detection performance degrades significantly for adversarial examples generated from OOD datasets except MNIST (Table 3). When  $\epsilon$  equals to 16, more than 45% adversarial examples generated from ImageNet, VOC12, and Internet Photographs datasets are missed by the detector.

However, in comparison to ODIN [44], the gradient-based attacks for this detector fail to achieve close to 100% FNR. We observe that even with an *unconstrained* adversarial attack, the FNR doesn't approach 100%. We speculate that this behavior might be due to non-informative gradients presented by the model at the input. It should be noted that first-order attack approaches which can succeed in presence of obfuscated gradients [3] aren't applicable here. This is because instead of any additional input-processing step the gradients are obfuscated by the model itself.

## 5.2 Fooling robustly trained models

In this section, we first evaluate the robustness of baseline, undefended models (trained with natural training) to OOD adversarial examples. Next, we evaluate models that are robustly trained using the two state-of-the-art approaches discussed in Section 2.3.2. This novel examination is critical because once the adversary manages to bypass the OOD detector, the open-world learning framework will pass that input to the relevant classifier. Our results highlight vulnerability of these approaches to OOD adversarial examples.

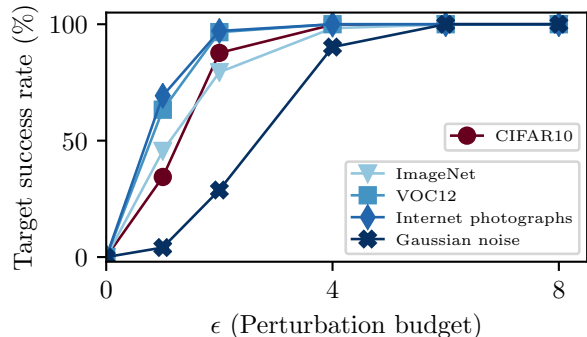
**5.2.1 Baseline models are highly vulnerable to OOD adversarial examples.** Fig. 2a shows the target success rate with adversarial examples generated from different OOD datasets for the Wide ResNet model trained on the CIFAR-10 dataset. The results show that similar to adversarial examples generated from in-distribution images, *OOD adversarial examples also achieve a high target success rate*. For example, the target success rate increases rapidly to 100% for both in- and out-of-distribution data.

**5.2.2 OOD attacks on adversarially trained models.** Previous work [26, 37, 46, 64] has shown that adversarial training can significantly increase the model robustness against adversarial examples generated from in-distribution data. For example, for the WRN-28-10 network trained on CIFAR-10, adversarial training reduces the target success rate from 100% to 22.9% for  $\epsilon$  equal to 8, (Figure 2b).

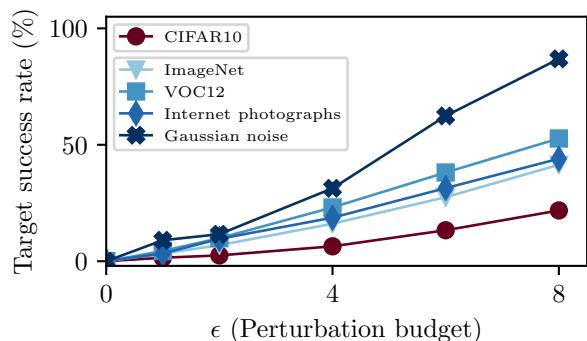
**Experimental details:** Models corresponding to MNIST, CIFAR-10 datasets in this experiment are  $M_1$ , WRN-28-10-A (Table 2) respectively. Each model is trained using iterative adversarial training [46] with an  $L_\infty$  perturbation budget ( $\epsilon$ ) equal to 0.3, 8 respectively.

**OOD adversarial examples generated from OOD datasets (except MNIST) achieve high target success rates for multiple adversarially trained models and datasets:** As highlighted in Fig. 2b for the CIFAR-10 classifier, although adversarial training improves robustness for in-distribution dataset (CIFAR-10), OOD adversarial examples achieve up to 4× higher target success rate compared to in-distribution adversarial examples.

Table 4 presents the detailed results for different datasets. We can see the improvement in target success rate with OOD adversarial examples. When using the VOC12 dataset to generate OOD adversarial examples, we can achieve around 66.7× and 2.4× improvement in target success rate compared to in-distribution attacks



(a) Lack of robustness of natural training.



(b) Lack of robustness of iterative adversarial training [46].

**Figure 2: Target success rate of adversarial examples generated from different datasets for the state-of-the-art WRN-28-10 [69] model trained on CIFAR-10 [39]. The PGD attack is used (Section 4.3) with  $\epsilon$  ( $l_\infty$  perturbation budget) up to 8. Though iterative adversarial training [46] (with  $\epsilon = 8$ ) improves robustness for in-distribution data (CIFAR-10), OOD adversarial examples are up to 4× as successful as those generated from in-distribution data.**

for MNIST and CIFAR-10 models respectively. The mean classification confidence for OOD adversarial examples is also competitive with adversarial examples generated from in-distribution data and typically higher than unmodified OOD data.

**5.2.3 OOD attacks on robust training with the convex polytope relaxation.** Robust training using convex polytope relaxation from Wong et al. [38, 65] provides a *provable* upper bound on the adversarial error and thus on target success rate.

**Experiment Details:** We use the models from Wong et al. [38, 65] for MNIST and CIFAR-10 dataset. The corresponding Models are  $M_2$  and WRN-28-1 respectively (Table 2). Given that this defense approach cannot simultaneously maintain high benign accuracy for CIFAR-10 while using a more realistic  $\epsilon$  equals to 8, we make a design choice of  $\epsilon$  equals to 2 in training the CIFAR-10 model and  $\epsilon$  equals to 8 for adversarial example generation. On the other hand, for simpler dataset such as MNIST, we continue to use the same  $\epsilon$  equals to 0.1 for both training and adversarial example generation.



**Table 4: Target success rate of adversarial examples generated from different datasets for models trained with iterative adversarial training [46], adversarial logit pairing (ALP) [37], and convex polytope relaxation [38]. The  $l_\infty$  norm used to generate adversarial examples is listed along with the training dataset. The maximum target success by the adversarial examples for every model is highlighted in bold. The results for in-distribution data are highlighted in *italics*.**

Test ( $\downarrow$ ) \ Train ( $\rightarrow$ ) dataset	Iterative adversarial training [46]						Adversarial logits pairing [37]			Provable Defenses [38]					
	MNIST ( $\epsilon = 0.3$ )			CIFAR-10 ( $\epsilon = 8$ )			ImageNet ( $\epsilon = 16$ )			MNIST ( $\epsilon = 0.1$ )			CIFAR-10 ( $\epsilon = 8$ )		
	success rate (%)	confidence clean	confidence adv	success rate (%)	confidence clean	confidence adv	success rate (%)	confidence clean	confidence adv	success rate (%)	confidence clean	confidence adv	success rate(%)	confidence clean	confidence adv
MNIST	<i>1.5</i>	<i>0.98</i>	<i>0.76</i>	5.1	0.81	0.60	98.8	0.27	0.96	<i>0.6</i>	<i>0.97</i>	<i>0.64</i>	3.9	0.48	0.37
CIFAR-10	97.6	0.77	0.99	22.9	<i>0.93</i>	<i>0.81</i>	<b>100.0</b>	0.14	0.99	67.2	0.88	0.97	<i>15.1</i>	<i>0.27</i>	<i>0.41</i>
ImageNet	97.2	0.79	0.99	44.9	0.74	0.78	<i>99.4</i>	<i>0.30</i>	<i>0.98</i>	72.1	0.88	0.97	23.4	0.40	0.36
VOC12	99.3	0.76	0.99	54.9	0.75	0.79	99.9	0.19	0.99	68.7	0.89	0.97	26.4	0.38	0.36
Internet Photographs	95.4	0.74	0.99	46.3	0.74	0.80	100.0	0.17	0.99	58.4	0.89	0.97	19.8	0.43	0.35
Gaussian Noise	<b>100.0</b>	0.92	1.00	<b>87.9</b>	0.52	0.86	48.5	0.41	0.45	<b>79.0</b>	0.91	0.99	<b>29.1</b>	0.31	0.32

**Robust training with convex polytope relaxation lacks robustness to OOD adversarial examples:** Although this defense significantly improves the robustness for in-distribution adversarial examples, it lacks robustness to OOD adversarial examples (Table 4). For the MNIST classifier, the target success rate increases from 0.6% to 72.1% by using ImageNet as a source of OOD data with  $\epsilon = 0.1$ . For the CIFAR-10 classifier, the target success rate increases from 15.1% to 29.1% with OOD adversarial examples generated from Gaussian noise. The relatively poor performance of adversarial examples for the CIFAR-10 classifier could be due to the *poor classification accuracy of this model*, where it achieves only 66.2% classification accuracy on the CIFAR-10 images. We argue that the principle behind this defense is not robust as demonstrated by success of OOD attacks on the provably trained MNIST model.

*5.2.4 Discussion: Impact of OOD dataset.* In this subsection, we further discuss the influence of dataset selection and robust learning on the success of evasion attacks. We observe in Table 4 that the target success rate is affected by the choice of the OOD dataset. In particular, we observe that the target success rate for MNIST dataset is significantly lower than both in-distribution and other OOD datasets. We speculate that this behavior could arise due to the specific semantic structure of MNIST images. Nevertheless, we emphasize that the threat posed by OOD adversarial examples still persists, since *adversarial examples from multiple other OOD datasets achieve high target success rates*.

### 5.3 Evading Adversarial Example Detectors and Secondary Defenses

*5.3.1 Adversarial example detectors.* Previous work [10, 29] has shown that both adversarial detectors based on Feature Squeezing [68] or MagNet [48] approach can be evaded with adaptive

white-box attacks accounting for the detector mechanism to generate adversarial examples from in-distribution data. Our results (in the Appendix) show that similar to in-distribution data, *these adversarial detectors don't provide robustness to adversarial examples generated from OOD data*. For feature squeezing, we also observe that OOD adversarial examples requires a smaller  $L_2$  perturbation budget than in-distribution adversarial attacks for a similar target success rate. We further show that OOD adversarial examples can achieve up to 97.3% target success rate in presence of MagNet on a CIFAR-10 classifier.

*5.3.2 Adversarial logit pairing.* Adversarial logit pairing [37] was the first technique to extend iterative adversarial training to ImageNet scale dataset. However, ALP suffers from loss of robustness for in-distribution adversarial examples when the number of attack iterations is increased [21]. We show that this vulnerability also exists for OOD adversarial examples (Table 4).

### 5.4 Towards real-world attacks

*5.4.1 Attacking Content Moderation systems.* To achieve low false positive rate (FPR), ML classifiers deployed in the real-world are also expected to detect OOD inputs. This is because a high FPR can significantly affect the performance [49] and cost [1] of the service provided by these models. For example, the London police are attempting to use computer vision to detect nudity in photographs, but a high FPR is occurring due to the prevalence of desert scenes as wallpapers etc. [49]. This example represents an inadvertent *denial-of-service (DoS) attack*, where a large number of false positives affects the effectiveness of the automated content moderation system. We use OOD adversarial examples to carry out a similar DoS attack on Clarifai's content moderation model [17], by *classifying clearly unobjectionable content as objectionable with high confidence*. In a real deployment, a deluge of such data

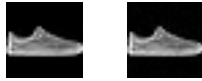


Figure 3: OOD adversarial examples against Clarifai’s Content Moderation model. Left: original image, classified as ‘safe’ with a confidence of 0.96. Right: adversarial example with  $\epsilon = 16$ , classified as ‘explicit’ with a confidence of 0.9.

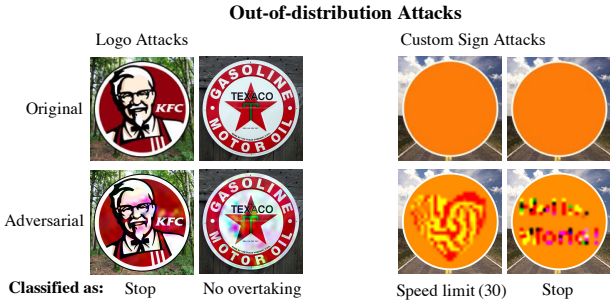


Figure 4: OOD adversarial examples for an open-world traffic-sign recognition pipeline. These adversarial examples are classified as the desired target traffic sign with high confidence under a variety of physical conditions when printed out.

will force human content moderators to spend time reviewing safe content. Sybil attacks [18] enable attackers to create fake accounts to upload large amounts of OOD adversarial examples.

**Fooling the Clarifai model:** Using the query-based black-box attacks proposed by Bhagoji et al. [7], we construct OOD adversarial examples for Clarifai’s content moderation model. It provides confidence scores for the input image belonging to the 5 classes ‘safe’, ‘suggestive’, ‘explicit’, ‘gore’ and ‘drugs’, and is accessible through an API. We use 10 images each from the MNIST, Fashion-MNIST [66] and Gaussian Noise datasets to generate OOD adversarial examples. All of these images are initially classified as ‘safe’ by the model. In Figure 3, we show a representative attack example. Our attack is able to successfully generate OOD adversarial examples for the 4 classes apart from ‘safe’ for all 30 images with 3000 queries on average and a mean target confidence of 0.7.

**5.4.2 Physically-realizable attacks on traffic signs.** We demonstrate success of physical OOD adversarial examples using both imperceptible perturbations in a OOD logo attack and unconstrained perturbations within a mask in a custom sign attack. OOD adversarial examples in Figure 4 are undetected and classified with high confidence as traffic signs (by an open-world system comprising a CNN with 98.5% accuracy on test data and the ODIN detector [44]) over a range of physical conditions when printed out. The targeted attack success rate is 95.2% for the custom sign attack. Further details and results are in the full version of this paper [58].

## 6 TOWARDS ROBUST OPEN-WORLD DEEP LEARNING

In this section, we present experimental results for our proposed hybrid combination of iterative adversarial training and selective

prediction to enhance classifier robustness against OOD adversarial examples. We experiment with image classification task with CIFAR-10 dataset as in-distribution and multiple other datasets as the source of out-of-distribution images. To train the background class, we include 5,000 out-of-distribution images along with 50,000 in-distribution images of CIFAR-10 dataset. Further details on experimental setup can be found in Appendix A.

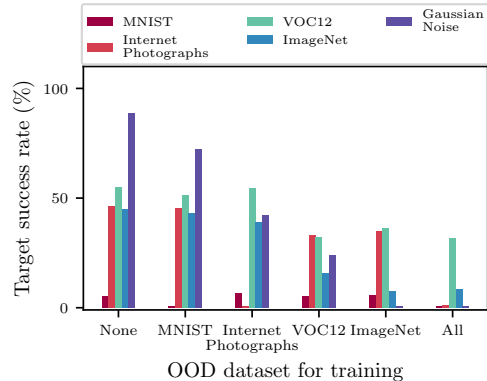


Figure 5: Target success rate of OOD adversarial examples (lower is better). The classifier is adversarially trained on in-distribution inputs along with a small subset of OOD data. It shows that datasets such as VOC12 and ImageNet provide a high inter-dataset and intra-dataset generalization for detection of adversarial OOD inputs.

**Small subset of OOD data can enable robust detection.** With each of the OOD dataset, we observe that using only 5,000 training images leads to significant decrease in the success of adversarial attacks from these datasets (Fig. 5). For example, after adversarial training with only 5,000 training images (out of 1.2 million) from ImageNet, the target success of OOD adversarial examples from ImageNet dataset decreases from 44.9% to 7.4%.

**Robust training with one OOD dataset can generalize to multiple other OOD datasets.** Figure 5 shows that adversarial training with one dataset decreases the attack success of OOD adversarial examples from other datasets. This effect is significant for feature-rich datasets such as ImageNet and VOC12, supported by the negative-bias property highlighted by Torralba et al. [63]. For example, using VOC12 for training reduces the target success rate of OOD adversarial examples from ImageNet from 44.9% to 15.8%. **Multiple OOD datasets can be combined for robust detection:** By including 5,000 images from each of the four OOD datasets in robust training, we demonstrate that a single network can learn robust detection for each of them. We observe that the combination of all datasets is constructive as the best results are achieved when multiple datasets are used in training.

**Small impact on in-distribution performance:** Robust training with OOD data has a small impact on classifier performance for in-distribution data. The maximum decrease in benign classification accuracy is 1% (for single OOD dataset) and 3.1% (for all four OOD datasets). The robust accuracy remains largely unchanged. **Discussion and limitations.** Our results highlight that it is feasible to robustly classify one or multiple OOD datasets along with

in-distribution data using a semi-supervised learning approach. However, the key challenge is to achieve robustness against all OOD inputs. As a first step, our approach and results motivate the design of robust unsupervised OOD detectors for deep learning. It was recently shown that using some unlabeled data during training can also improve robustness against in-distribution adversarial attacks [12, 60]. Our results also highlight the need for rigorous evaluation methods to determine the robustness of open-world learning systems against all possible adversarial examples.

## 7 CONCLUSION

In this paper, we investigated evasion attacks in the open-world learning framework and defined OOD adversarial examples, which represent a new attack vector on machine learning models used in practice. We found that existing OOD detectors are insufficient to deal with this threat. Analyzing the robustness of alternative OOD detection mechanisms, such as autoencoders [14], and distance comparisons in feature space [36, 43], would be an interesting direction of future work.

Further, assumptions regarding the source of adversarial examples, namely, in-distribution data, have led to tailored defenses. We showed that these state-of-the-art defenses exhibit increased vulnerability to OOD adversarial examples, which makes their deployment challenging. With these findings in mind, we took a first step at countering OOD adversarial examples using adversarial training with background class augmented classifiers. We now urge the community to consider the exploration of strong defenses against open-world evasion attacks.

## ACKNOWLEDGMENTS

This research was sponsored by the National Science Foundation under grants CNS-1553437, CNS1704105, CIF-1617286 and EARS-1642962, by Intel through the Intel Faculty Research Award, by the Office of Naval Research through the Young Investigator Program (YIP) Award and by the Army Research Office through the Young Investigator Program (YIP) Award. ANB would like to thank Siemens for supporting him through the FutureMakers Fellowship.

## REFERENCES

- [1] 2008. <https://www.absolute.com/en/go/reports/the-cost-of-insecure-endpoints>. (2008). [Online; accessed 10-November-2018].
- [2] 2019. Pictum Random image generator. "<https://picsum.photos/>". (2019).
- [3] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*.
- [4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing Robust Adversarial Examples. In *ICML*. 284–293.
- [5] Abhijit Bendale and Terrance E Boult. 2015. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1893–1902.
- [6] Abhijit Bendale and Terrance E Boult. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1563–1572.
- [7] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. 2018. Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms. In *ECCV*.
- [8] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrđić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 387–402.
- [9] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Müller, Jiakai Zhang, et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).
- [10] Nicholas Carlini and David Wagner. 2017. MagNet and “Efficient Defenses Against Adversarial Attacks” are Not Robust to Adversarial Examples. *arXiv preprint arXiv:1711.08478* (2017).
- [11] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. 39–57.
- [12] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. 2019. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736* (2019).
- [13] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep Learning for Anomaly Detection: A Survey. *arXiv e-prints* (Jan. 2019). arXiv:cs.LG/1901.03407
- [14] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. 2017. Robust, deep and inductive anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 36–51.
- [15] Eric I Chang and Richard P Lippmann. 1994. Figure of merit training for detection and spotting. In *NeurIPS*. 1019–1026.
- [16] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*. 2722–2730.
- [17] Clarifai 2019. Clarifai | Image & Video Recognition API. <https://clarifai.com>. (2019).
- [18] George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting Sybil Nodes using Social Networks.. In *NDSS*. San Diego, CA, 1–15.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
- [20] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. 2018. Reducing Network Agnostophobia. In *NeurIPS*. 9175–9186.
- [21] Logan Engstrom, Andrew Ilyas, and Anish Athalye. 2018. Evaluating and Understanding the Robustness of Adversarial Logit Pairing. *arXiv preprint arXiv:1807.10272* (2018).
- [22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.
- [23] Ivan Evtimov, Kevin Eykholt, Earlene Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2018. Robust Physical-World Attacks on Machine Learning Models. In *IEEE conference on computer vision and pattern recognition*.
- [24] Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *NeurIPS*. 4878–4887.
- [25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT Press.
- [26] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.
- [27] Manuel Günther, Steve Cruz, Ethan M Rudd, and Terrance E Boult. 2017. Toward open-set face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE.
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [29] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. 2017. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*.
- [30] D. Hendrycks and K. Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.
- [31] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2019. Deep Anomaly Detection with Outlier Exposure. In *ICLR*.
- [32] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR abs/1704.04861* (2017).
- [33] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [34] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and Artificial Intelligence*. ACM, 43–58.
- [35] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box Adversarial Attacks with Limited Queries and Information. In *ICML*. 2142–2151.
- [36] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. In *NeurIPS*. 5546–5557.
- [37] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. 2018. Adversarial Logit Pairing. *arXiv preprint arXiv:1803.06373* (2018).
- [38] J Zico Kolter and Eric Wong. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*.
- [39] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2014. The CIFAR-10 dataset. *online: http://www.cs.toronto.edu/kriz/cifar.html* (2014).
- [40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NeurIPS*. 1097–1105.

- [41] Yann LeCun. 1998. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/> (1998).
- [42] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *ICLR*.
- [43] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*. 7167–7177.
- [44] S. Liang, Y. Li, and R. Srikant. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR*.
- [45] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.
- [46] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.
- [47] Michael McCoyd and David Wagner. 2018. Background Class Defense Against Adversarial Examples. In *2018 IEEE Security and Privacy Workshops (SPW)*. 96–102.
- [48] Dongyu Meng and Hao Chen. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 135–147.
- [49] Margi Murphy. 2017. Artificial intelligence will detect child abuse images to save police from trauma. <https://www.telegraph.co.uk/technology/2017/12/18/artificial-intelligence-will-detect-child-abuse-images-save/>. (2017).
- [50] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 427–436.
- [51] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep face recognition.. In *BMVC*, Vol. 1. 6.
- [52] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. In *ICLR*.
- [53] Manikandasriram Srinivasan Ramanagopal, Cyrus Anderson, Ram Vasudevan, and Matthew Johnson-Roberson. 2018. Failing to learn: autonomously identifying perception failures for self-driving cars. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3860–3867.
- [54] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In *ICLR*.
- [55] Vikash Sehwal, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. 2019. Better the Devil you Know: An Analysis of Evasion Attacks using Out-of-Distribution Adversarial Examples. *arXiv preprint arXiv:1905.01726* (2019).
- [56] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1528–1540.
- [57] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR* (2015).
- [58] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Prateek Mittal, and Mung Chiang. 2018. Rogue Signs: Deceiving Traffic Sign Recognition with Malicious Ads and Logos. In *DLS (IEEE SP)*.
- [59] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2015. Striving for simplicity: The all convolutional net. *ICLR* (2015).
- [60] Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. 2019. Are Labels Required for Improving Adversarial Robustness? *arXiv:1905.13725* (2019).
- [61] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. Deepid3: Face recognition with very deep neural networks. *arXiv:1502.00873* (2015).
- [62] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.
- [63] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1521–1528.
- [64] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*.
- [65] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. 2018. Scaling provable adversarial defenses. *NeurIPS* (2018).
- [66] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
- [67] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. 2019. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 501–509.
- [68] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *NDSS*.
- [69] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *BMVC*.

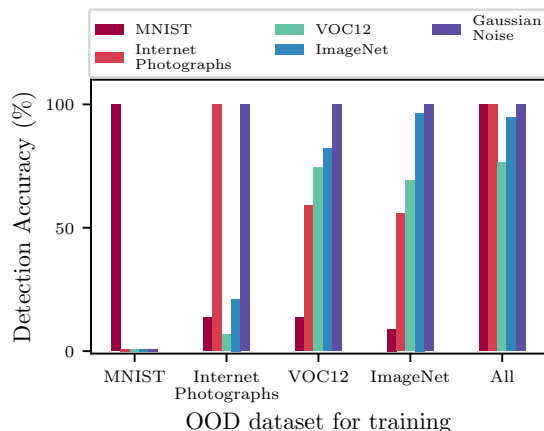
## A ADDITIONAL RESULTS

**Table 5: Target success rate for adversarial examples with random target labels from different datasets in presence of adversarial detectors, including feature squeezing [68] and MagNet [48]. Similar to in-distribution adversarial examples, OOD adversarial examples are also able to evade the adversarial detectors with a high success rate.**

Test (↓) \ Train (→) dataset	Feature squeezing [68]			MagNet [48]	
	MNIST	CIFAR-10	ImageNet	MNIST ( $\epsilon = 0.3$ )	CIFAR-10 ( $\epsilon = 8$ )
MNIST	98.1	100.0	3.12	<b>32.3</b>	18.0
CIFAR-10	99.2	100.0	96.1	0.7	90.1
ImageNet	99.3	100.0	85.1	0.8	92.5
VOC12	99.3	<b>100.0</b>	96.0	0.8	96.9
Internet	98.1	100.0	85.1	1.2	<b>97.3</b>
Photographs					
Gaussian	<b>100.0</b>	100.0	25.0	0.0	93.9
Noise					

**Adversarial Example Detectors:** We present the results for OOD attacks (Table 5) against the state-of-art adversarial example detectors feature squeezing [68] and MagNet [48] (recall Section 2.3.3). Target success rate in presence of these detectors refers to the percentage of adversarial examples which both evade the detectors and achieve target label after classification.

**Robust open-world machine learning:** To train the classifier in presence of background class, we use 5,000 images from one of the datasets from MNIST, ImageNet, VOC12, and Random Photographs. The reason to include only 5,000 images is to avoid data bias when each class in the CIFAR-10 dataset has 5,000 images. To include multiple OOD datasets, we add one background class for each. Figure 6 represent the success of the classifier in rejecting the out-of-distribution inputs. It shows that datasets such as VOC12 and ImageNet provide a high inter-dataset and intra-dataset generalization for detection of unmodified OOD inputs.



**Figure 6: Classification accuracy of unmodified OOD inputs, which is the percentage of OOD inputs classified to the background class.**