

Towards Characterizing Adversarial Defects of Deep Learning Software from the Lens of Uncertainty

Xiyue Zhang*
Peking University,
China

Xiaofei Xie*
Nanyang Technological University,
Singapore

Lei Ma†
Kyushu University,
Japan

Xiaoning Du
Nanyang Technological University,
Singapore

Qiang Hu
Kyushu University,
Japan

Yang Liu
Nanyang Technological University,
Singapore

Jianjun Zhao
Kyushu University,
Japan

Meng Sun†
Peking University,
China

ABSTRACT

Over the past decade, deep learning (DL) has been successfully applied to many industrial domain-specific tasks. However, the current state-of-the-art DL software still suffers from quality issues, which raises great concern especially in the context of safety- and security-critical scenarios. Adversarial examples (AEs) represent a typical and important type of defects needed to be urgently addressed, on which a DL software makes incorrect decisions. Such defects occur through either **intentional attack or physical-world noise perceived by input sensors**, potentially hindering further industry deployment. The intrinsic uncertainty nature of deep learning decisions can be a fundamental reason for its incorrect behavior. Although some testing, adversarial attack and defense techniques have been recently proposed, **it still lacks a systematic study to uncover the relationship between AEs and DL uncertainty**.

In this paper, we conduct a large-scale study towards bridging this gap. We first investigate the capability of multiple uncertainty metrics in differentiating benign examples (BEs) and AEs, which enables to characterize the uncertainty patterns of input data. Then, we identify and categorize the uncertainty patterns of BEs and AEs, and find that while BEs and AEs generated by existing methods do follow common uncertainty patterns, some other uncertainty patterns are largely missed. Based on this, we propose an automated testing technique to generate multiple types of uncommon AEs and BEs that are largely missed by existing techniques. Our further evaluation reveals that the uncommon data generated by our

method is hard to be defended by the existing defense techniques with the average defense success rate reduced by 35%. Our results call for attention and necessity to generate more diverse data for evaluating quality assurance solutions of DL software.

CCS CONCEPTS

• **Software and its engineering** → **Software testing and debugging**; • **Computing methodologies** → *Neural networks*.

KEYWORDS

Deep learning, uncertainty, adversarial attack, software testing

ACM Reference Format:

Xiyue Zhang, Xiaofei Xie, Lei Ma, Xiaoning Du, Qiang Hu, Yang Liu, Jianjun Zhao, and Meng Sun. 2020. Towards Characterizing Adversarial Defects of Deep Learning Software from the Lens of Uncertainty. In *42nd International Conference on Software Engineering (ICSE '20)*, May 23–29, 2020, Seoul, Republic of Korea. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3377811.3380368>

1 INTRODUCTION

In company with the booming of available domain-specific big data and hardware acceleration, deep learning (DL) experienced big performance leap in the past few years, in achieving competitive performance in many cutting edge applications (e.g., image processing [45], speech recognition [19], sentiment analysis [6], e-commerce recommendation [59], video game control [36]). However, the state-of-the-art DL software still suffers from quality issues. A deep neural network (DNN) that achieves high prediction accuracy can still be vulnerable to adversarial examples (AEs) [2]. For example, an image recognition DL software can be easily fooled by pixel-level noises [51] or noises perceived in a physical-world situation [8, 52]. The quality and reliability issues, without properly addressing or confining, could potentially hinder more widespread adoption of DL software especially in the applications with higher requirements of safety and security (e.g., autonomous driving, medical diagnosis).

The incorrect decision of a DL software can trace back to several typical sources and patterns (e.g., generalization capability issue, robustness issue [21]). Up to present, AEs remain to be one

* Co-first authors

† Lei Ma (malei@ait.kyushu-u.ac.jp) and Meng Sun (sunm@pku.edu.cn) are the corresponding authors.

Permission to make digital or hard copies of all or part 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 components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICSE '20, May 23–29, 2020, Seoul, Republic of Korea

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7121-6/20/05...\$15.00

<https://doi.org/10.1145/3377811.3380368>

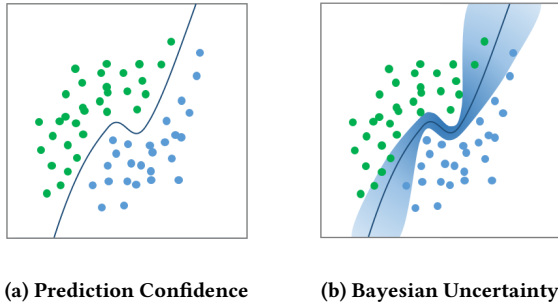


Figure 1: Prediction Confidence and Bayesian Uncertainty.

of the most notable types of DL defects, which reveals the quality and robustness issues of the DL software. Although the arms races of many recently proposed adversarial attack [4, 14, 24, 37], defense [13, 15, 17, 40, 43, 53, 57] and testing [7, 41, 52, 55] continuously escalate, most of these techniques are rather ad-hoc. Thus far, research efforts on understanding and interpretation of AEs and benign examples (BEs) are still at an early stage [50]. Uncertainty provides a new perspective to characterize AEs and BEs, towards understanding and designing better quality assurance techniques for DL software, e.g., testing, adversarial attack/defense.

To bridge this gap, in this paper, we study 4 state-of-the-art Bayesian uncertainty metrics (based on the statistical analysis of multi-shot executions) along with the one-shot execution metrics, i.e., prediction confidence score (PCS) of the DL software under analysis (see Fig. 1 and definitions in § 2). We perform a large-scale comparative study to investigate the sensitivity/capability of these metrics in differentiating AEs and BEs, which are important indicators to characterize DL runtime behaviors. We find that PCS and Variation Ratio in terms of original prediction (VRO) (§ 2) are among the best candidates with such capability, and they are selected for further input data characterization. Then, an obvious question arises: *what is the relation between the AEs/BEs generated by the state-of-the-art adversarial attack/testing techniques and these uncertainty metrics?* In particular, *do these AEs/BEs follow some patterns in terms of uncertainty?* Our in-depth analysis reveals that AEs/BEs generated by existing techniques [4, 14, 25, 37, 38, 55] largely fall into two common patterns (that we refer as common input samples): (1) AEs tend to have low PCS and high VRO uncertainty, and (2) BEs often come with high PCS and low VRO.

Based on the above observation, further questions naturally arise: *What do those uncommon inputs look like and can we generate them possibly and automatically?* and *Would these uncommon data inputs differ from common inputs, e.g., are they even more challenging to be correctly handled by a DL software?* To answer these questions, we propose a genetic algorithm (GA) based automated test generation technique that iteratively generates uncommon input samples guided by uncertainty metrics. We implement the proposed technique as a tool named *KuK* (to Know the UnKnown), and demonstrate its effectiveness in generating uncommon inputs on a large benchmark, including three datasets MNIST, CIFAR10, ImageNet across four different DL model architectures *LeNet-5*, *NIN*,

ResNet-20, *MobileNet*. In line with existing adversarial defense techniques, our comparative experiments against the state-of-the-art adversarial attack/testing techniques also reveal that the uncommon samples could be more challenging to be correctly handled by DL software in many cases. Such uncommon samples represent a new type of hazard and potential defects to DL software, which so far lacks investigation and should draw further attention during future quality assurance solution design.

In summary, this paper investigates the following research questions with the support of a large-scale study:

- **RQ1:** What is the capability of the state-of-the-art uncertainty metrics in differentiating AEs and BEs?
- **RQ2:** Do the AEs and BEs generated by the state-of-the-art adversarial attack/testing techniques follow common uncertainty patterns? If so, what are such common patterns?
- **RQ3:** Is it possible to generate those uncommon data that is missed by the state-of-the-art attack/testing techniques? Is *KuK* useful in generating uncommon data?
- **RQ4:** To what extent are the uncommon samples defended by existing adversarial defense techniques compared with the common ones?

Through answering RQ1 and RQ2, we aim to characterize behaviors of DL software on AEs and BEs from the uncertainty perspective, and investigate whether some common patterns are followed by data inputs generated through current state-of-the-art adversarial attack/testing techniques. Our study confirms the existence of the common patterns for such generated data (i.e., low PCS and high VRO for AEs, high PCS and low VRO for BEs). It also identifies new uncommon sample categories which are missed by existing techniques.

RQ3 and RQ4 focus on understanding the feasibility to obtain the uncommon samples (i.e., RQ3) and the impact of such samples on the quality and reliability of DL software (i.e., RQ4). Our evaluation results confirm that the uncommon samples could be generated with proper testing guidance. Such uncommon samples, representing a new type of generated test data, could bypass a variety of adversarial defense techniques with higher success rates. Thus, it is quite important to generate such uncommon inputs to reveal the hidden defects of DL software for vulnerability analysis and further enhancement, especially for safety- and security-critical scenarios. We believe such uncommon data could be an important clue towards building trustworthy DL solutions, which should draw special attention for further quality assurance solution design.

The contributions of this paper are summarized as follows:

- We perform an empirical study on four state-of-the-art Bayesian uncertainty metrics and one prediction confidence metric to investigate their ability to differentiate BEs and AEs, i.e., data inputs that can/cannot be correctly handled by a DL software. Among these metrics, PCS and VRO outperform the others in achieving higher differentiating accuracy.
- We perform a systematic study on the BEs and AEs generated by six state-of-the-art adversarial attack/testing techniques (i.e., FGSM [14], BIM [25], Deepfool [37], C&W [4], DeepHunter [55], TensorFuzz [38]) to identify potential uncertainty patterns in terms of PCS and VRO. Our results reveal that (1) AEs and BEs from existing techniques largely follow two common patterns

while the other patterns (denoted as uncommon patterns) are largely missed by existing methods, and (2) the characteristics of AEs generated by the testing tools differ from those by the adversarial attack techniques.

- We propose a GA-based automated testing technique for DL software, implemented as a tool *KuK*, towards generating input samples with diverse uncertainty patterns. Especially, our evaluation on three datasets MNIST, CIFAR10, ImageNet across four different model architectures *LeNet-5*, *NIN*, *ResNet-20*, *MobileNet* demonstrates its effectiveness in generating uncommon input samples that are largely missed by existing techniques.
- We further investigate how the adversarial defense techniques [13, 15, 17, 40, 43, 53, 57] react to the uncommon samples, in line with the samples generated by adversarial attack techniques. Our results indicate that the current defense techniques are often biased towards particular patterns of samples. The generated uncommon samples can bypass such defense techniques with a high success rate, potentially causing severe threats to the quality and reliability of DL software. For example, on the model *NIN*, the uncommon data achieve 97.5% success rate on bypassing the mutation-based defense, while the common data only make it 5.5%.

2 PRELIMINARIES AND OVERVIEW

2.1 Deep Neural Networks

DEFINITION 1 (DEEP NEURAL NETWORK). A deep neural network (DNN) M is a function $P = M(x)$ that maps an input x to a predictive probability vector P . The output label of M on input data x is $L_M(x) = \arg\max_{i \in C} P[i]$ where C is the set of classes and $P = M(x)$.

In general, a DNN learns to extract features from the distribution of training data layer by layer, and provides the decision on each candidate class with some probabilistic confidence. A higher predictive probability value of a class often indicates higher prediction confidence on that decision label.

DEFINITION 2 (BENIGN EXAMPLE). A benign example (BE) x with ground truth label y is an input sample, such that the prediction decision of a DNN M is consistent with the ground truth label: $L_M(x) = y$.

Benign examples refer to those inputs that could be correctly handled by a given DNN model M .

DEFINITION 3 (ADVERSARIAL EXAMPLE). An adversarial example (AE) is an input x' similar to a benign example x by adding some minor perturbation δ , (i.e., $x' = x + \delta$), but resulting in a different prediction decision of a DNN M (i.e., $L_M(x') \neq L_M(x)$).

Existing attack methods usually generate the AEs by manipulating the output of the *logit* layer or *softmax* layer of a DNN M , which gradually decreases in the output probability of ground truth label and increases in the probability of other labels. In this way, the prediction decision is shifted to other labels. Based on direct observations on the prediction results of such AEs, we identify that a typical type of unreliable prediction is usually accompanied by two classes that have close probability confidence values.

DEFINITION 4 (PREDICTION CONFIDENCE SCORE). Given a DNN M and an input x , the prediction confidence score (PCS) of the input

x on M is defined as

$$PCS(x, M) = \max_{i \in C} P[i] - \max_{i \in C \setminus c^*} P[i]$$

where C is the set of classes, $c^* = L_M(x)$ and $P = M(x)$.

Intuitively, PCS depicts the probability difference between the two classes with the highest probabilities, which provides an uncertainty proxy from the aspect of distance to the geometric boundary. For an input x , the smaller the $PCS(x, M)$, the closer x is to the decision boundary between the top-two classes. As a result, it is more likely to cross the boundary with noise perturbations. In the following sections, we use $PCS(X, M)$ to denote $\{PCS(x, M) | x \in X\}$, where X is a set of inputs.

2.2 Bayesian Uncertainty Measures

Besides prediction confidence based metrics, Bayesian-based methods are recently proposed to estimate DNN's uncertainty through multi-shot execution analysis [11] (see Fig. 1). From the principled Bayesian perspective, DNNs are not regarded as deterministic functions with fixed parameters. Instead, the parameters of DNNs are treated as random variables, which obey a *prior* distribution $p(\omega)$. The posterior distribution $q(\omega | \mathcal{D})$ is then approximated given a training data set \mathcal{D} , based on which uncertainty estimates can be obtained. The relationship between the uncertainty representative of *prediction confidence score* and *Bayesian uncertainty estimates* is shown in Fig. 1. Intuitively, prediction confidence score reflects a distance abstraction to the fixed geometric boundary w.r.t. a point-estimate neural network whose parameters are fixed; while Bayesian method ensembles a set of networks whose weights follow some probability distribution.

However, due to the complexity of DNNs, it is often impractical to sample infinitely many weights from the distribution to perform the runtime execution. To this end, the state-of-the-art approach to obtain uncertainty estimates makes use of the dropout technique from multiple runs (*Monte Carlo dropout* [12]). Although dropout is originally proposed as a regularization method in the training process to avoid over-fitting, in the context of uncertainty estimation, dropout is leveraged in the testing process, which samples weights from the distribution to obtain DNN instances. As a result, it ensembles a number of neural networks with different weights. In each prediction execution, it randomly drops out some units in the DNN, which may cause different prediction results. As a result, it allows to obtain uncertainty estimates efficiently and scales to real-world neural networks. Specifically, there are three commonly-used metrics to estimate uncertainty [11], i.e., variation ratio, predictive entropy and mutual information.

2.2.1 Variation Ratio. Variation ratio measures the dispersion from the dominant class of the prediction (i.e., the predicted class with the highest frequency in multiple predictions).

DEFINITION 5 (VARIATION RATIO). Given a model M and an input x , the variation ratio (VR) of the input x is defined as

$$VR(x, M) = 1 - \frac{\sum_{k \in \{1, \dots, T\}} L_{M_d}^k(x) = l_{max}}{T}$$

where M_d is the model with dropout-enable and $L_{M_d}^k$ denotes the k -th prediction result by M_d . T is the total number of prediction

execution by M_d . l_{max} represents the dominant label from most of the T predictions.

Another variant of the *variation ratio* in terms of the original prediction (the prediction of the model under analysis) is defined as follows:

DEFINITION 6. Given a model M and an input x , the *variation ratio for original prediction* (denoted as VRO) of the input x is defined as

$$VRO(x, M) = 1 - \frac{\sum_{k \in \{1, \dots, T\}} L_{M_d^k}(x) = L_M(x)}{T}$$

where $L_M(x)$ is the prediction result from the original model M .

Intuitively, VR measures the general uncertainty of the decision with the highest frequency (i.e., whether most predictions agree with the same result) while VRO represents the stability around the prediction mode of model M (i.e., whether the majority predictions agree with the original result). The higher the VR or VRO is, the more uncertain the prediction is.

2.2.2 Predictive Entropy. Predictive entropy originates from information theory and measures the average amount of information contained in a stochastic source of predictive output. When all classes are predicted with equal probability in the form of a uniform distribution, the decision carries the most information, indicating high uncertainty. In contrast, when one of the classes is predicted with high probability value (e.g., 0.9), then the prediction is relatively certain.

DEFINITION 7. Given all the predictive probability distributions $P_d^k = M_d^k(x)$, $k \in \{1, \dots, T\}$ of an input x across T predictions on dropout enabling model M_d , the *predictive entropy* is defined as

$$PE(x, M) = - \sum_{i \in C} \left(\frac{1}{T} \sum_k P_d^k(i) \right) \log \left(\frac{1}{T} \sum_k P_d^k(i) \right)$$

where $P_d^k(i)$ denotes the probability value of a particular class i on the k -th prediction of model M_d .

2.2.3 Mutual Information. Mutual information quantifies the amount of information obtained about one random variable through the observations of the other random variable. In the case of DNNs, the two random variables are prediction $L_M(x)$ and the posterior of the model parameters ω , whose distribution is approximated through T stochastic forward passes of dropout enabled model M_d . We use ω_k to denote an instance of the model parameters sampled from the posterior distribution.

DEFINITION 8. Given the predictive probability distributions $P_d^k = M_d^k(x)$, $k \in \{1, \dots, T\}$ of an input x in T predictions, the *mutual information of input x* is defined as

$$MI(x, M) = PE(x, M) + \frac{1}{T} \sum_{i,k} (P_d^k(i) \log P_d^k(i))$$

where $P_d^k(i)$ is used to approximate $P_d(i|\omega_k)$ in a similar way to predictive entropy, based on the probability vectors obtained through different stochastic forward passes.

Table 1: Subject datasets and DNN models.

Dataset	DNN Model	#Neuron	#Layer	Acc.(%)
MNIST	LeNet-5	268	9	99.0%
CIFAR-10	NIN	1418	12	88.2%
	ResNet-20	2,570	70	86.9%
ImageNet	MobileNet	38,904	87	87.1%*

* The reported top-5 test accuracy of pretrained DNN model in [20].

2.3 Overview

In this paper, we aim to understand the capability of different uncertainty portrayals, on which behaviors of AEs and BEs are further characterized. Fig. 2 shows an overview of our work, summarized as two major components: (1) an empirical study about the uncertainty metrics and characteristics of the existing data inputs, and (2) the data generation algorithm that generates input samples with uncommon patterns and potential applications. Specifically, we first perform an empirical study towards understanding the capability of different uncertainty metrics on distinguishing AEs and BEs (§ 3.2). Then, we propose a way of categorizing AEs and BEs based on two perspectives: the *prediction confidence score* from the perspective of the single-shot model execution and the *uncertainty estimates* from the statistical perspective of multi-shot execution. Next, we study the uncertainty patterns of the BEs and AEs generated by the existing adversarial attack/testing tools (§ 3.4).

From the empirical study, we find that the existing AEs and BEs follow specific uncertainty patterns. Then, We propose a genetic-algorithm-based approach to generate the uncommon data inputs (§ 4), whose uncertainty patterns are different from the patterns that existing data fall into. We identify and analyze the importance of data with diverse uncertainty patterns from a testing perspective. Finally, we evaluate the capability of such uncommon data in bypassing a variety of defense techniques (§ 5.2).

3 EMPIRICAL STUDY

In this section, we first perform a comparative study about the capability of different uncertainty metrics in distinguishing AEs and BEs. Then, we conducted a follow-up investigation of the characteristics of existing AEs/BEs from the uncertainty perspective.

3.1 Subject Dataset and Data Preparation

3.1.1 Datasets. We selected three popular publicly available datasets (i.e., MNIST [26], CIFAR-10 [23], and ImageNet [45]) as the evaluation subject datasets (see Table 1).

MNIST is for handwritten digit image recognition, containing 60,000 training data and 10,000 test data, with a total number of 70,000 data categorized in 10 classes (i.e., handwritten digits from 0 to 9). Each MNIST image is of size $28 \times 28 \times 1$ with a single channel.

CIFAR-10 is a collection of images for general-purpose image classification, including 50,000 training data and 10,000 test data in 10 different classes (e.g., airplane, bird, cat and ship), with 6,000 images per class. Each CIFAR-10 image is of size $32 \times 32 \times 3$ with three channels. The classification task of CIFAR-10 is generally difficult than that of MNIST due to the data size and complexity.

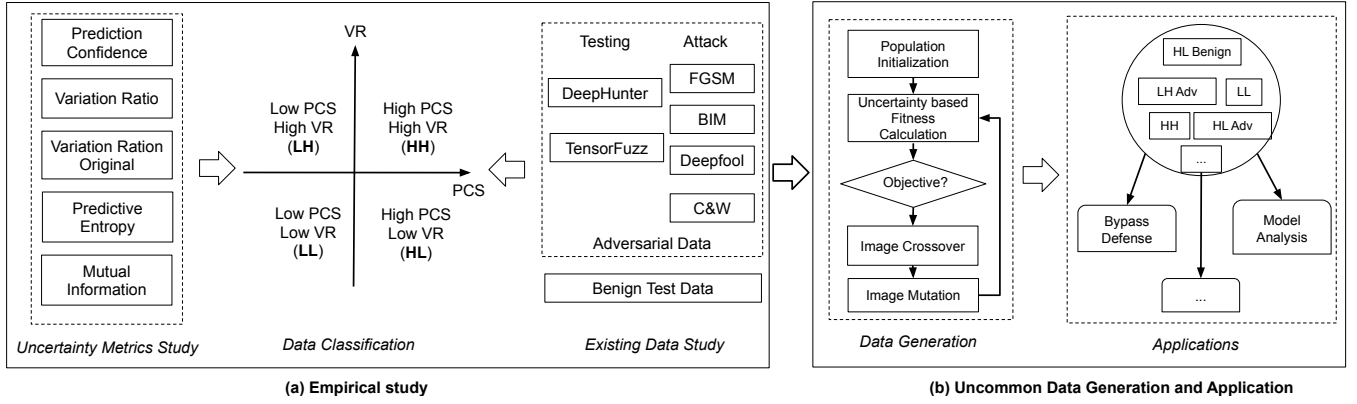


Figure 2: The overview of our study and its application.

ImageNet is a large-scale practice-sized dataset, which is used as the database for large-scale visual recognition challenge (ILSVRC) towards general-purpose image classification. The complexity of *ImageNet* is characterized by a training set comprised of over 1 million images, together with a validation set comprised of 50,000 images and a test set with 150,000 images. Each image is of size $224 \times 224 \times 3$.

For each dataset, we study several popular DNN models used in previous work [18, 20, 26, 27], which achieve competitive test accuracy (Table 1).

The DL models we study in this paper are all with convolutional architectures. However, our approach is generic and could be applied to other network architectures such as recurrent neural networks. Our approach focuses on the uncertainty nature of DNN. Whether the calculation methods of the uncertainty metrics are available determines the feasibility of applying our approach to DL models with other architectures. Basically, the only requirement on the DL models is with a classification setting, thus the model output would be a probability distribution among a set of classes and the uncertainty metrics can be obtained. In a word, our approach can be applied to DNN classification tasks, independent of the model architectures.

3.1.2 Adversarial Example Generation Tools. We chose four state-of-the-art adversarial attack tools, *i.e.*, FGSM [14] (Fast Gradient Sign Method), BIM [25] (Basic Iterative Method), Deepfool [37] and C&W [4] attacks, to generate adversarial examples. Specifically, we used the existing Python tool-set FOOLBOX [44] to perform these attacks and each attack is configured with the default setting.

We also selected two state-of-the-art automated testing tools for deep neuron networks, *i.e.*, DeepHunter [55] and TensorFuzz [38], both of which adopt coverage-based fuzzing techniques. For DeepHunter, we generated AEs with the k -Multisection Neuron Coverage (KMNC) and Neuron Coverage (NC) as the guidance. Following the configuration in [41, 55], we set the parameter k as 1,000 for KMNC and the threshold as 0.75 for NC. TensorFuzz is configured with the default setting used in [38].

3.1.3 Data Preparation. Overall, we prepared the following three sets of data: one set of benign examples, one set of AEs generated

Table 2: Number of adversarial examples generated by the testing tools and adversarial attacks.

Model	DH-KMNC	DH-NC	TensorFuzz	Adv_attacks
<i>LeNet-5</i>	2,980	4,607	1,436	4×9,000
<i>NIN</i>	5,715	9,110	3,848	4×9,000
<i>ResNet-20</i>	6,960	9,249	2,465	4×9,000
<i>MobileNet</i>	3,314	20,541	13,999	4×9,000

by the attack methods, and one set of AEs generated by testing tools (see Table 2):

- *BenignData*. For each dataset, we randomly sampled 9,000 test data, which can be correctly predicted by the models, as the benign dataset for each model.
- *AttackAdv*. For each input in *BenignData*, we generated four types of AEs with the four attack methods, resulting in a total of 36,000 AEs. (Column *Adv_attacks* in Table 2).
- *TestAdv*. For each dataset, we randomly sampled 500 benign examples as the initial seeds. Then, we ran DeepHunter and TensorFuzz for each model with 5,000 iterations. The columns *DH-KMNC*, *DH-NC*, and *TensorFuzz* show the number of adversarial examples generated by DeepHunter with KMNC and NC guidance, and Tensorfuzz.

3.2 RQ1: Empirical Study on Uncertainty Metrics

The objective of RQ1 is to study the relationship between uncertainty metrics and adversarial examples. In particular, we analyze the effectiveness of uncertainty metrics in distinguishing AEs and BEs. We adopt AUC-ROC [9] to evaluate the classification performance of each metric. We utilize the score as the evaluation criterion because it measures the performance without dependence on a pre-setting threshold. Specifically, AUC-ROC is a performance measurement for classification problems at various threshold settings. ROC is short for receiver operating characteristic curve and AUC represents degree or measure of separability. It gives us detailed information on to what extent the evaluated model is capable

of distinguishing between classes. In our metric effectiveness evaluation, the higher the AUC-ROC score, the better a metric is at distinguishing AEs and BEs.

Table 3 shows the AUC-ROC scores achieved on different metrics across different data. To be specific, we used 9,000 BEs and 9,000 AEs generated from each attack to calculate the AUC-ROC scores. Overall, PCS achieves the best performance as it is a direct measure of the prediction confidence of the target model. Interestingly, we found that AEs often have low prediction confidence.

From the Bayesian uncertainty perspective, we found that VRO achieves the best performance. On MNIST and CIFAR-10 datasets, all AUC-ROC scores are over 97%. On ImageNet, the minimum AUC-ROC score is 77.47% in differentiating BEs and AEs generated by Deepfool, while the scores on other attacks are all above 84%. From the in-depth analysis, we found that VRO (refer to Definition 6) captures the difference between the prediction of the original model and multiple predictions of the randomized dropout-enabled model. On the other hand, other metrics represent the uncertainty based on the stability of the multi-shot predictions without considering the model under analysis. For example, given an input, suppose the prediction of the original model is incorrect and all predictions of the dropout-enabled model are correct, the values of VR and VRO would be 0 and 1, respectively. Intuitively, VRO shows that the model is quite uncertain but VR indicates that the model is very certain. In other words, VRO is more sensitive to capture the uncertain behavior of the target model compared with other uncertainty metrics.

Answer to RQ1: PCS of the original model is often effective in distinguishing BEs and AEs generated by existing attacks. For the Bayesian uncertainty metrics, VRO is often more effective than others when comparing the prediction stability between the original model and multiple dropout-enabled predictions.

3.3 Characterizing Data Behavior

PCS and Bayesian uncertainty depict the prediction results of a DNN from different angles. In particular, existing studies demonstrate that high prediction confidence is not equal to low Bayesian uncertainty, and vice versa [11].

The prediction confidence (*i.e.*, PCS) represents the confidence of a single-shot model execution while the Bayesian uncertainty is measured by the statistical results from multi-shot model executions. From RQ1, the results reveal that VRO stands out to capture the different behaviors of BEs and AEs compared with other Bayesian uncertainty metrics. Therefore, we adopt a two-dimensional metric, *i.e.*, (PCS, VRO), to characterize the BEs and AEs on a specific DNN model. Based on this, the data is classified into four patterns (see Fig. 2 (a)): low PCS and high VRO (LH), high PCS and high VRO (HH), low PCS and low VRO (LL) and high PCS and low VRO (HL). The categorization provides a way to understand and analyze the behaviors of AEs and BEs.

3.4 RQ2: Categorization of Existing Data

Based on the categorization method, we perform a study towards understanding the characteristics of BEs and AEs generated by

Table 3: The AUC-ROC scores of classification models with different metrics.

Model	Attacks	PCS	VRO	VR	PE	MI
LeNet-5	BIM	99.98%	99.06%	90.72%	83.56%	81.52%
	C&W	100.00%	99.08%	90.23%	82.86%	81.61%
	Deepfool	99.44%	98.31%	93.47%	86.46%	84.78%
	FGSM	99.98%	98.74%	95.52%	90.09%	87.36%
NIN	BIM	99.95%	99.46%	88.57%	86.73%	86.95%
	C&W	99.90%	99.44%	87.93%	85.99%	86.48%
	Deepfool	99.79%	99.18%	91.44%	88.64%	88.51%
	FGSM	99.43%	98.80%	93.97%	91.69%	91.54%
ResNet-20	BIM	99.97%	98.20%	86.74%	84.62%	85.10%
	C&W	99.88%	98.28%	85.87%	83.93%	84.59%
	Deepfool	99.80%	97.85%	88.02%	85.70%	86.23%
	FGSM	99.23%	97.28%	90.74%	88.25%	88.24%
MobileNet	BIM	99.95%	86.67%	84.25%	68.77%	68.37%
	C&W	96.80%	84.49%	82.73%	67.80%	67.35%
	Deepfool	79.36%	77.47%	77.13%	74.03%	72.32%
	FGSM	97.79%	87.93%	86.49%	72.75%	71.84%

adversarial attacks and testing tools. To compute the VRO, we follow the parameter configuration suggested in [12] and set T (see Definition 6) as 50 for MNIST, and 100 for CIFAR-10 and ImageNet, respectively.

Table 4 summarizes the quantitative results of BEs and AEs, w.r.t. the two-dimensional metrics in four models. Note that the data in the columns *Benign*, *AEs from Attacks* and *AEs from Testing tools* are from the three sets of data presented in § 3.1.3. In each cell, the two values represent the mean and variance of the corresponding PCS and VRO metric results, respectively.

Overall, benign data mostly have high PCS and low VRO. The mean values of PCS for the four models are 0.99, 0.953, 0.945 and 0.788, respectively, while the mean values of VRO are 0.293, 0.054, 0.072 and 0.337. The results are mostly in line with our expectations, *i.e.*, BEs are expected to be predicted by the model with high confidence and low uncertainty. In the case of MobileNet, the PCS is relatively smaller than others while the VRO is larger than that of other models. The reason is that MobileNet handles a more complex task, *i.e.*, image classification for a large-scale dataset (ImageNet). It is more difficult to train a high-quality DNN model. MobileNet we used here is among the state-of-the-art model for image classification in ImageNet, whose top-5 accuracy is 87.1%. This result indicates that BEs usually belong to the HL type.

For AEs generated by different attacks, the metric performance is mostly contrary to BEs, *i.e.*, AEs generated by attacks usually come with low PCS and high VRO. Except for the AEs generated by Deepfool on ImageNet, all the mean values of PCS are rather low (*e.g.*, almost all PCS values are below 0.1) and the mean values of VRO are relatively high (*e.g.*, all VRO values are larger than 0.652). It indicates that AEs generated from state-of-the-art adversarial attacks usually fall into the category with low confidence and high uncertainty. From the results, we find that AEs usually belong to the LH type.

For the adversarial data generated from testing tools, we found that the obtained metrics are between BEs and AEs generated from the adversarial attacks. The mean values of PCS are smaller than

Table 4: Results (mean / variance) of benign & adversarial data.

Model	Metric	Benign	AEs from Attacks				AEs from Testing tools		
			BIM	C&W	Deepfool	FGSM	DH-KMNC	DH-NC	TensorFuzz
<i>LeNet-5</i>	PCS	0.990 / 0.004	0.018 / 0.001	0.002 / 8.865e-06	0.186 / 0.114	0.012 / 1e-4	0.561/0.119	0.592/0.112	0.579/0.111
	VRO	0.312 / 0.017	0.733 / 0.008	0.735 / 0.008	0.697 / 0.009	0.716 / 0.009	0.631/0.019	0.625/0.020	0.630/0.020
<i>NIN</i>	PCS	0.953 / 0.023	0.007 / 0.0004	0.013 / 0.0003	0.027 / 0.007	0.083 / 0.005	0.571/0.113	0.608/0.110	0.384/ 0.093
	VRO	0.054 / 0.016	0.777 / 0.025	0.777 / 0.027	0.720 / 0.026	0.679 / 0.029	0.380/0.065	0.352/0.063	0.615/0.039
<i>ResNet-20</i>	PCS	0.945 / 0.026	0.005 / 0.0002	0.016 / 0.0005	0.023 / 0.006	0.096 / 0.006	0.548/0.108	0.574/0.109	0.392/0.093
	VRO	0.072 / 0.023	0.865 / 0.010	0.869 / 0.010	0.850 / 0.010	0.828 / 0.012	0.398/0.080	0.388/0.080	0.457/0.062
<i>MobileNet</i>	PCS	0.788 / 0.082	0.002 / 1.598e-05	0.085 / 0.020	0.415 / 0.135	0.059 / 0.002	0.601/0.160	0.659/0.119	0.614/0.126
	VRO	0.337 / 0.062	0.679 / 0.015	0.652 / 0.017	0.596 / 0.054	0.696 / 0.016	0.454/0.080	0.451/0.077	0.551/0.083

the PCS results of BEs but often larger than the results of AEs from attacks. Conversely, the mean values of VRO are larger than the VRO results of BEs but smaller than the results of AEs from adversarial attacks. For example, consider the metric results of BEs, TensorFuzz AEs and C&W AEs on ResNet-20, the PCS values are 0.945, 0.392 and 0.016, respectively, while the VRO values are 0.072, 0.457 and 0.869. Compared with the results of BEs, we can still reach a similar conclusion that AEs from testing tools also belong to the LH type.

Even so, the results of AEs from adversarial attacks and testing tools have some differences. This might be caused by their differences in the test generation methods. Current adversarial attacks usually adopt the gradient-based or optimization-based technique to gradually decrease the predictive probability or the logit value of the truth label until the decision result changes. For example, when the probability of the truth label is reduced to 0.49 and the probability of another label becomes 0.51, an adversarial example is found and the attack stops. DeepHunter and TensorFuzz adopt the mutation-based technique to generate new tests. The random mutation can not guarantee that the predictive probability of the truth label is decreased gradually. For example, the probability may change from 0.99 to 0.10 by only one mutation, resulting in a higher PCS. In summary, although both adversarial attacks and testing tools can generate AEs, the behaviors of such generated AEs are different in terms of PCS and VRO. Therefore, this further confirms the difference between testing and adversarial attack, from the perspective of software engineering. Testing can not only simulate real-world scenarios to uncover potential issues of a DL software for deployment, but also generate more diverse data to capture the behaviors of the DNN systematically in the applied context.

Comparing the results of BEs and AEs, we also found that there also exists a tentative *inverse correlation* between PCS and VRO. For example, the PCS of BEs is often large and the VRO is small. The PCS of AEs from adversarial attacks is often small while the VRO is large. For AEs generated from testing tools, the PCS tends to be larger and the VRO is smaller. It is reasonable since high prediction confidence usually reflects that the prediction is relatively certain.

Answer to RQ2: BEs and AEs usually belong to the HL type and LH type, respectively. Compared with state-of-the-art adversarial attacks, testing tools to some extent generate different AEs.

4 UNCOMMON DATA GENERATION

The results of RQ2 reveal that BEs and AEs usually belong to the HL type and LH type, respectively. However, several questions still remain, *i.e.* whether there exist: 1) data samples with high PCS and high VRO (*i.e.*, the HH type), 2) data with low PCS and low VRO (*i.e.*, the LL type), 3) BEs with low PCS and high VRO (*i.e.*, the LH BEs) and 4) AEs with high PCS and low VRO (*i.e.*, the HL AEs). These samples have the potential to uncover the unknown behaviors of DNN, which are largely missed by existing methods. To answer these questions, we developed a tool, *KuK*, to generate such uncommon data.¹

As PCS and VRO on existing data usually follow an inverse correlation, it is non-trivial to generate the uncommon data. In fact, the test data generation of specific types could be a complicated optimization problem. In this paper, we leverage the Genetic Algorithm (GA) [35] to provide a solution.

Fig. 2 b) shows the workflow of our algorithm. The inputs of *KuK* include a seed x (*i.e.*, an initial image²), a model M , the *dropout* enabled model M_d and a target type c . The output is a set of data samples that satisfy the *objective*. We elaborate on the details of each step as follows.

Population Initialization. Given an input image, we first generate a set of images by randomly adding noise to it. In order to generate high-quality images (*i.e.*, recognizable by human), we abandon the affine transformation (*e.g.*, translation and rotation [55]) as the crossover may generate invalid images. We use L_∞ norm to constrain the allowable changes between the original seed and its generated counterpart.

Objective and Fitness Calculation. In each iteration, we check whether some generated images (in the population) satisfy the *objective*, which is specifically designed for each type based on PCS and VRO. The test generation continues until some desirable data inputs are obtained. To satisfy the objective, we design a set of *piecewise fitness functions* to generate different types of uncommon data such that the higher the corresponding fitness value, the better the input.

We use X to denote the population, *i.e.*, a set of images, and use $PCS(X, M)$ to denote $\{PCS(x, M) | x \in X\}$. Given the input x and

¹We refer to these data as "uncommon" in the sense that they are rarely uncovered by existing techniques rather than that they rarely exist. Such data could occur widely in the real world.

²In this paper, although we mainly focus on the image domain, the approach can also generalize to other domains.

model M , the objectives and the fitness functions are defined as follows:

- For the *LL* type, the objective is $PCS(x, M) < p \wedge VRO(x, M) < v$, where p and v are configurable parameters, and the fitness function is:

$$Fit_{LL}(x) = \begin{cases} -PCS(x, M), & \min(PCS(X, M)) > p \\ (PCS(x, M) < p) - VRO(x, M), & \text{otherwise} \end{cases} \quad (1)$$

If the minimum PCS of the population is larger than p , we use $-PCS(x, M)$ to decrease the PCS until there are some inputs whose PCSs are below p . Due to the custom inverse correlation between PCS and VRO, VRO tends to increase when PCS decreases. As for the fitness function in the other situation, $(PCS(x, M) < p)$ aims to ensure that the PCS is still below p while $-VRO(x, M)$ stands out for a smaller VRO.

- For the *HH* type, the objective is $PCS(x, M) > p \wedge VRO(x, M) > v$, where p and v are configurable parameters, and the fitness function is:

$$Fit_{HH}(x) = \begin{cases} PCS(x, M), & \max(PCS(X, M)) < p \\ (PCS(x, M) > p) + VRO(x, M), & \text{otherwise} \end{cases} \quad (2)$$

If the maximum PCS of X is smaller than p , we increase the PCS until some PCSs are larger than p . In addition, we ensure the PCSs is larger than p ($PCS(x, M) > p$), and in the meantime attempt to increase VRO ($+VRO(x, M)$).

- To generate AEs which belong to the *HL* type, the objective is set as: $PCS(x, M) > p \wedge VRO(x, M) < v \wedge x$ is an AE, where p and v are the configurable parameters, and the fitness function is:

$$Fit_{HL}^A(x) = \begin{cases} -PCS(x, M), & \forall x' \in X, x' \text{ is BE} \\ AE(x) + PCS(x, M), & \exists x' \in X, x' \text{ is AE} \\ & \wedge \max(PCS(X, M)) < p \\ AE(x) + (PCS(x, M) > p) & \text{otherwise} \\ -VRO(x, M), & \end{cases} \quad (3)$$

where $AE(x)$ is 1 if x is an AE. Otherwise, it is 0.

The generation of HL AEs is extremely challenging since HL is the typical feature of BEs. To address the problem, we design a three-step approach. If all images in the population are BEs (step 1), we aim to generate AEs by decreasing the PCS, which is commonly used by state-of-the-art attacks. Whenever any inputs become AEs and all PCSs become smaller than p (step 2), we increase PCS but still keep the high priority of AE (*i.e.*, AE has a high fitness value with the support of $AE(x)$). For example, if an AE becomes BE but achieves high PCS, its fitness value will decrease. In the last step, we set AE and high PCS as the high priority, then decrease VRO.

- To generate BEs that belong to *LH* type, the objective is set as: $PCS(x, M) < p \wedge VRO(x, M) > v$ and x is a BE. The fitness function is designed as follows:

$$Fit_{LH}^B(x) = \begin{cases} BE(x) - PCS(x, M), & \min(PCS(X, M)) > p \\ BE(x) + (PCS(x, M) < p) & \text{otherwise} \\ +VR(x, M) & \end{cases} \quad (4)$$

where $BE(x)$ is 1 if x is a BE. Otherwise, it is 0.

Similarly, the generation of LH BEs is also challenging as LH is a typical feature of AEs. In the first step, if all PCSs are larger than p , we first decrease the PCS but keep the high priority of BE. In

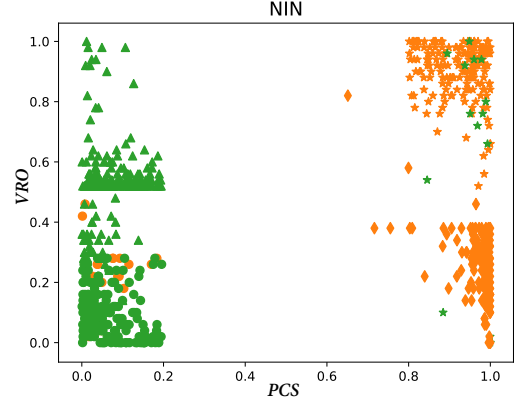


Figure 3: The distribution of different types of data generated on NIN by KuK. The orange color represents AEs, while green color represents BEs. Star, circle, triangle and diamond represent the data of HH, LL, LH and HL, respectively.

the second step, when there are some BEs with low PCSs, we increase the VRO but keep the high priority of BE and low PCS.

Crossover and Mutation. For the crossover, we adopt the tournament selection strategy to select two tournaments. From each tournament, we select one image, which has the largest fitness value. The two selected images are used to perform the crossover by randomly exchanging the corresponding pixels. After the crossover, each image is randomly mutated by adding white noise, to increase the diversity of the population. The test generation continues until the objective is satisfied or the given computation resource (*e.g.*, time limit) exhausts.

5 EVALUATION

We implemented the proposed test generation tool, *KuK*, in Python based on Keras [5] (2.2.4) with TensorFlow [1] (1.12.0) as backend. In this section, we aim to evaluate the usefulness of *KuK* in generating uncommon data (RQ3) and the effectiveness of these data in bypassing the defense techniques (RQ4). All the experiments were run on a server with the Ubuntu 16.04 system with 28-core 2.0GHz Xeon CPU, 196 GB RAM and 4 NVIDIA Tesla V100 16G GPUs.

5.1 RQ3: Usefulness of Test Data Generation

Setting. We adopt *KuK* to generate different types of uncommon data on four widely used DL models – *LeNet-5*, *NIN*, *ResNet-20* and *MobileNet*. In the genetic algorithm, the size of the population is set as 100, the crossover rate is set as 0.5 and the mutation rate is 0.005. For the mutation process, the radius of L_∞ is set as 0.3. For each dataset, we randomly select 200 BEs as the initial seeds. For each seed, we generate four types of uncommon data. The maximum number of the iterations in the genetic algorithm is set to 50.

Threshold Selection. To perform the categorization, we need to set the upper bound for the low PCS/VRO, the lower bound for the high PCS/VRO, *i.e.*, the configuration values of the parameters p and v in the objective and fitness functions. Actually, in Table 4, the results of BEs (HL) and AEs (LH) generated from adversarial

attacks are the extreme cases, which can act as the guidance to select the thresholds.

- *High PCS*. For the lower bound of high PCS, we set it as 0.7 as all high PCSs from BEs in the four models are above 0.7. Specifically, the minimum value of high PCS of BEs is 0.788 in *MobileNet*. Intuitively, if the PCS of a data sample is above 0.7, we regard it as a data sample with *high PCS*.
- *Low VRO*. For the upper bound of the low VRO, we set the value as 0.4 since the low VROs of BEs in the four models are below 0.4 (e.g., 0.312 and 0.337 for *LeNet-5* and *MobileNet*, respectively). If the VRO of a data sample is below 0.4, we categorize it into *low VRO* type.
- *Low PCS*. For the upper bound of the low PCS, most of the low PCSs of AEs in Table 4 are below 0.1 while the PCS of samples generated by Deepfool for *MobileNet* has a larger value 0.415. We made a compromise and set the value as 0.3. If the PCS of a data sample is below 0.3, we regard that it falls into *low PCS* category.
- *High VRO*. For the lower bound of high VRO, we set it as 0.6 because almost all of the high VROs of AEs in the four models are larger than 0.6 (one exception is 0.596 in *MobileNet*). If the VRO is above 0.6, we regard it as *high VRO*.

Results. Fig. 3 depicts the distribution of the 200 generated data on the two-dimension plane (due to page limit, the results of other models are put on our website [54]). Note that there are some seeds from which we failed to generate the uncommon data satisfying the objective. For each of these seeds, we plot the best result from the population (e.g., for HH type, we select the data, which has the maximum of the sum of PCS and VRO). The results show that *KuK* enables to generate inputs with diverse uncertainty patterns.

Table 5 shows the number of uncommon data inputs that satisfy the objective and the mean PCS and VRO value. Row *Type Objective* shows the objective setting for each uncommon uncertainty pattern in terms of upper and lower bound of PCS and VRO. For each model, Row *Total* and *PCS/VRO* give the total number of uncommon data generated for each type and the mean value of PCS and VRO, respectively. For LL and HH types, the results of generated AEs and BEs are shown separately (i.e., Column *Ben* and Column *Adv*).

The results demonstrate that *KuK* is effective in generating LL and HH data inputs that are rarely covered by existing methods. For example, for *NIN* model, *KuK* generated 198 (99%) LL data in total, of which 176 data are BEs and 22 data are AEs. For *ResNet-20*, *KuK* generated the LL data and HH data for all seeds (i.e., 200). From the quantitative result of LL data, we could find that LL data tend to be BEs (e.g., the number of LL BEs is much larger than the number of LL AEs). Considering the natural BEs usually belong to HL (refer to Table 4), it indicates that low VRO is a better metric to represent the characteristics of BEs. For the HH data, there is no such obvious trend. In particular, for *LeNet-5* and *MobileNet*, the number of HH BEs is larger than the number of HH AEs. However, the case in *NIN* and *ResNet-20* is on the contrary.

For LH BEs and HL AEs, it is more challenging to generate them since they are completely opposite to the characteristics of the common data. The results show that *KuK* is useful to generate such uncommon data. For example, we generated 172 (86%), 58 (29%), 67 (33.5%) and 152 (76%) LH BEs for *LeNet-5*, *NIN*, *ResNet-20* and *MobileNet*, respectively. For HL AEs, we found that *KuK* only

generated one HL AE for *LeNet-5*. For other models, *KuK* generated 165 (82.5%), 93 (46.5%) and 74 (37%) HL AEs, respectively. The results indicate that generating LH BEs and HL AEs is more difficult and *KuK* can still generate them for a part of seeds.

We can see that the PCS and VRO in Table 5 are consistent with those in Table 4. For example, although we set the lower bound of high PCS as 0.7 in the fitness functions and objectives, *KuK* still generated very high PCS for HH and HL data. The mean value of PCS is larger than 0.9 in *LeNet-5*, *NIN*, and *ResNet-20*, which is very consistent with the high PCS of BEs in Table 4. Even we set the upper bound of low VRO as 0.4, *KuK* still generated data with pretty low VRO, e.g., for *NIN* and *ResNet-20*, the VROs of LL data are 0.095 and 0.105, respectively.

Comparing the results among the four models, we could find that the difficulty in generating uncommon data varies for different models. For example, *KuK* is effective in generating LL and HH data for *LeNet-5*, *NIN*, and *ResNet-20*, but only generated 88 (44%) and 81 (40.5%) for *MobileNet*. For *LeNet-5* and *MobileNet*, it is easier to generate LH BEs than HL AEs but the case in *NIN* and *ResNet-20* is on the contrary. These differences show that the uncommon data generated through *KuK* can be used to characterize different behaviors of the models.

Answer to RQ3: *KuK* is useful for generating different types of uncommon data. The HL BEs and HL AEs are often more difficult to be generated.

5.2 RQ4: Evaluation on Defense Techniques

To demonstrate the usefulness of the generated uncommon data in Table 5, this experiment intends to study whether the data can bypass the existing defense techniques.

Setting. Since different defense techniques are proposed on different subject datasets, we selected popular techniques based on the datasets. For MNIST and CIFAR10 dataset, we selected the following defense techniques: binary activation classifier [13], mutation-based adversarial attack detection [53], defensive distillation [40], label smoothing [17], and feature squeezing [57]. For ImageNet, we selected the mutation-based adversarial attack detection [53], input transformations [15] and pixel deflection [43]. Due to the space limit, we put the details about the configuration and the introduction of each defense technique on our website [54].

To validate the performance of the defense techniques, we selected 1) the common data including 9,000 BEs and 9,000 AEs generated from the existing adversarial attacks (§ 3.1.3), and 2) the uncommon data from Table 5. We use the success rate to evaluate the capability of the defense technique, i.e., divide the number of BEs/AEs, which can be correctly identified, by the total number.

Results. Table 6 and 7 show the results about the performance of the defense techniques on four models. Column/Row *Comm* represents the success rate on existing common data. The results show that the defense techniques are very effective in identifying BEs and AEs of the common data, especially on the smaller models. For example, except for that the success rate of feature squeezing is 88.1% on *ResNet-20*, the success rates of other defense techniques are above 90% on *NIN*, *ResNet-20*, and *LeNet-5*. In particular, the success rate is above 97% on *LeNet-5*. For the larger model *MobileNet*, the

Table 5: Results of different types of data generated by KuK.

Type Objective		LL (<0.3/<0.4)		HH (>0.7/>0.6)		LH (<0.3/>0.6)	HL (>0.7/<0.4)
		Ben	Adv	Ben	Adv	Ben	Adv
LeNet-5	Total	124	18	130	46	172	1
	PCS/VRO	0.067/0.324	0.030/0.343	0.947/0.720	0.903/0.803	0.082/0.609	0.99/0.339
NIN	Total	176	22	10	184	58	165
	PCS/VRO	0.065/0.095	0.065/0.204	0.960/0.846	0.922/0.910	0.073/0.666	0.981/0.245
ResNet-20	Total	168	32	11	181	67	93
	PCS/VRO	0.065/0.105	0.064/0.224	0.960/0.851	0.921/0.910	0.058/0.655	0.978/0.253
MobileNet	Total	83	5	70	11	152	74
	PCS/VRO	0.234/0.348	0.192/0.348	0.845/0.714	0.784/0.785	0.092/0.719	0.915/0.328

success rate is relatively low as it is more difficult to perform the defense for the complex model.

Column/Row *UnCo* represents the success rate on the uncommon data generated from *KuK*. The overall result shows that the existing defense techniques perform poorly on the uncommon data we generated. We could find that the success rate of the *binary classifier* and *mutation-based* detection are reduced a lot. For example, on *NIN*, the success rate is reduced to 25.5% and 5.5%, respectively. The reason is that these two techniques mainly depend on the PCS and VRO characteristics for detection. Specifically, *binary classifier* is trained with the value of *logit* layer that is closely related to prediction confidence and the label change ratio in *mutation-based* detection is similar to VRO. The uncommon data is very different from the common data *w.r.t.* these two metrics. As a result, if they perform well on common data, the success rate on the uncommon data could be low.

For other defense techniques, the reduction in success rate appears smaller than that of *binary classifier* and *mutation-based detection*. For example, the success rates drop to 78.3% and 76.2% for *defensive distillation* and *label smoothing* on *NIN*. The reason is that a new model is retrained with these defense techniques, while the attacks are generated regarding to the original one, making it a more challenging transfer attack scenario. For example, *defensive distillation* retrain a more robust model by reducing the gradients. In this case, some of the data, which are uncommon for original model, become common data *w.r.t.* the retrained model, because of some weight variation. However, it still can be seen from the results that the uncommon data reveal stronger transferability.

We also found that the uncommon data on *LeNet-5* is not effective on *defensive distillation* and *label smoothing*. We performed an investigation and found that it may be caused by the following reasons. 1) Most of the data become common in the new model. It confirms the usefulness of uncommon data in characterizing the different behaviors of multiple models. 2) Most of the uncommon data generated on *LeNet-5* are BEs (see Table 5). The success rate is reduced to 90.8% and 79.9% if we only use the uncommon AEs.

Answer to RQ4: The uncommon data inputs are not well defended by existing defense techniques while the common data are relatively easier to be defended. In particular, the *binary classifier* and *mutation-based detection* approaches are less useful in defending the uncommon data inputs (e.g., with only 5% and 10% on *NIN* and *ResNet-20*, respectively).

Table 6: Success rate of the defense techniques on the generated data for *NIN*, *ResNet-20* and *LeNet-5*.

	NIN		ResNet-20		LeNet5	
	Comm	UnCo	Comm	UnCo	Comm	UnCo
#Data	18,000	615	18,000	552	18,000	491
binary classifier	0.944	0.255	0.958	0.159	0.986	0.303
mutation-based	0.975	0.055	0.985	0.101	0.97	0.390
distillation	0.93	0.786	0.913	0.773	0.985	0.963
label smoothing	0.936	0.762	0.921	0.748	0.981	0.934
feature squeezing	0.905	0.663	0.881	0.637	0.973	0.327

6 THREAT TO VALIDITY

The selection of the subject datasets and DNN models could be a threat to validity. We try to counter this by using three publically available datasets with diverse scales, and popular pre-trained DNN models that achieve competitive performance.

The selection of the thresholds for the categorization may affect the results of Table 5. We carefully select the thresholds based on the results of Table 4. Furthermore, the results of Table 5 are relatively consistent with the results of Table 4, indicating that the selection basically does not affect the results of Table 5.

A further threat would be the randomness factors for computing the VRO (*i.e.*, configuration parameter *T* in § 6). The previous work [3] found that the result is not sensitive to the choice of *T* as long as *T* is greater than 20 in CIFAR-10 and MNIST. We follow the configuration in the existing paper [12], *i.e.*, 50, 100, 100 for MNIST, CIFAR-10 and ImageNet. In addition, we tested that the values are sufficient for computing the stable VRO.

7 RELATED WORK

In this section, we summarize the most relevant work to ours.

Attack and defense. Ever since the demonstration of deep learning models being vulnerable to even a small perturbation of input data [49], a sequence of attack techniques have been developed on the strand. To date, multiple types of attacks including FGSM [14], JSMA [39], BIM [24], DeepFool [37], C&W [4] have been proposed; a parallel research focus on improving the robustness of deep learning models. Goodfellow et al. [14] presented a method of introducing nonlinear model families into the training process. Defensive distillation was introduced to reduce the effectiveness of adversarial samples [40], which was then broken by C&W attack [4]. Meanwhile, a set of recent defense techniques was surveyed and shown

Table 7: Success rate of the defense techniques on the generated data for *MobileNet*.

	#Data	mutation-based	input trans	deflection
Comm	18,000	0.865	0.702	0.782
UnCo	395	0.549	0.660	0.688

that all could be defeated by constructing new loss functions [3]. A more recent work [34] exploited the framework of robust optimization for network adversarial training to resist a wide range of attacks. Besides dealing with datasets like MNIST and CIFAR10, defense techniques [15, 43] were also proposed to handle real-world large-scale datasets like ImageNet. However, it still lacks a study on the characteristics of BEs and AEs generated through these methods, which we attempt to attain from the uncertainty perspective.

Uncertainty measures. In general, a deep learning model is trained with a dataset and results in a set of fixed parameters, which further sets up a deterministic function mapping an input to a probability distribution. Bayesian approach, however, does not view deep learning models as deterministic functions; instead, they treat the parameters as random variables [33]. As a representative work to solve the scalability problem of obtaining uncertainty measures, [12] proposed a dropout-based solution, which allows us to calculate uncertainty estimates of existing deep learning models with a good trade-off between uncertainty quality and computational complexity. Existing research on uncertainty measure applications mainly focuses on the adversary detection problem. For example, [10] used both density estimates and Bayesian uncertainty estimates to learn a regression model for adversarial example detection. Further, an empirical study on two types of uncertainty measures, predictive entropy and mutual information, was proposed to understand the effectiveness of them for detection [46]. However, in our work, we perform a comparative study on both single-shot and multi-shot execution uncertainty estimates to dig out the uncertainty patterns that existing AEs/BEs followed.

Testing and debugging. Researchers have attempted to leverage decades of advances in the software engineering community to seek for solutions towards more secure and robust DL systems and have developed a set of fruitful results. These approaches share a different spirit from those in the DL community, and, for the first time, have been evidenced to be unique and promising without studies. Testing criteria come out as the first research focus. A series of measurements have been adapted to evaluate the quality of DL test dataset, including neuron coverage [41], multi-granularity coverage criteria [29], MC/DC test criteria [47], combinatorial testing criteria [28], surprise adequacy [22] and uncertainty-based metrics [32]. Classic testing methodologies have also been incorporated for DNN testing, including differential testing [41], coverage-guided testing [38, 52, 55], mutation testing [30] and concolic testing [48]. Some advanced test generation methods [7, 56, 58] have also been proposed to achieve better testing for different applications. Similar to samples generated by adversarial attacks, it still lacks a study on the relationship of samples generated by testing techniques and uncertainty. In addition, the results of this paper confirm the difference between testing and adversarial attack, in obtaining samples with different uncertainty behaviors.

Some recent efforts have been made to debug DL models [31], and to study DL program bugs [60], library bugs [42] and DL software bugs across different frameworks and platforms [16]. The results of this paper provide a new angle to characterize DL model defects, which could be useful for other quality assurance activities besides testing.

8 CONCLUSION AND FUTURE WORK

This paper performed an empirical study to characterize the data inputs from the perspective of uncertainty. We first presented an empirical study on the capability of uncertainty metrics in differentiating AEs and BEs. Then, we performed a systematic study of the characteristics of BEs and AEs generated by existing attack/testing methods in terms of uncertainty metrics. The results reveal that existing BEs and AEs largely fall into two uncertainty patterns in terms of PCS and VRO. Based on the investigation results, we proposed a GA-based automated test generation technique to generate data with more diverse uncertainty patterns, especially those uncommon samples. The results demonstrated the usefulness of the generated data in bypassing defense techniques. In future, we plan to perform more in-depth investigations on the application of the uncommon data towards robustness enhancement. We believe further understanding of these uncommon data is crucial for building reliable and trustworthy DL solutions.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comprehensive feedback. This research was supported (in part) by the National Research Foundation, Prime Ministers Office, Singapore under its National Cybersecurity R&D Program (Award No.NRF2018NCR-NCR005-0001), National Satellite of Excellence in Trustworthy Software System (Award No.NRF2018NCR-NSOE003-0001); JSPS KAKENHI Grant No.19K24348, 19H04086, 18H04097, Qdai-jump Research Program No.01277; the National Natural Science Foundation of China under grant No.61772038, 61532019, and the Guangdong Science and Technology Department (Grant No.2018B010107004). We also gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC) to our research.

REFERENCES

- [1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 265–283.
- [2] 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.
- [3] Nicholas Carlini and David Wagner. 2017. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISeC '17)*. ACM, 3–14.
- [4] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*. 39–57.
- [5] François Chollet et al. 2015. Keras. <https://github.com/fchollet/keras>. (2015).
- [6] Hai Ha Dohaiha, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2018. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems With Applications* (2018).
- [7] Xiaoning Du, Xiaofei Xie, Yi Li, Lei Ma, Yang Liu, and Jianjun Zhao. 2019. DeepStellar: model-based quantitative analysis of stateful deep learning systems. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 477–487.
- [8] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1625–1634.
- [9] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.
- [10] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. 2017. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410* (2017).
- [11] Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph.D. Dissertation. University of Cambridge.
- [12] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. 1050–1059.
- [13] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. 2017. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960* (2017).
- [14] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*. <http://arxiv.org/abs/1412.6572>
- [15] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* (2017).
- [16] Qianyu Guo, Sen Chen, Xiaofei Xie, Lei Ma, Qiang Hu, Hongtao Liu, Yang Liu, Jianjun Zhao, and Xiaohong Li. 2019. An empirical study towards characterizing deep learning development and deployment across different frameworks and platforms. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 810–822.
- [17] Tamir Hazan, George Papandreou, and Daniel Tarlow. 2016. *Perturbations, Optimization, and Statistics*. MIT Press.
- [18] 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.
- [19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (Nov 2012), 82–97.
- [20] 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. *arXiv preprint arXiv:1704.04861* (2017).
- [21] Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. 2017. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24–28, 2017, Proceedings, Part I*. 97–117.
- [22] Jinhan Kim, Robert Feldt, and Shin Yoo. 2019. Guiding Deep Learning System Testing Using Surprise Adequacy. In *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*. 1039–1049.
- [23] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. [n. d.]. CIFAR-10 (Canadian Institute for Advanced Research). ([n. d.]). <http://www.cs.toronto.edu/~kriz/cifar.html>
- [24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).
- [25] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial Machine Learning at Scale. *arXiv preprint arXiv:1611.01236* (2016).
- [26] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [27] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).
- [28] Lei Ma, Felix Juefei-Xu, Minhui Xue, Bo Li, Li Li, Yang Liu, and Jianjun Zhao. 2019. DeepCT: Tomographic Combinatorial Testing for Deep Learning Systems. In *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. 614–618.
- [29] Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang. 2018. DeepGauge: Multi-granularity Testing Criteria for Deep Learning Systems. In *Proc. of the 33rd ACM/IEEE Intl. Conf. on Automated Software Engineering (ASE 2018)*. 120–131.
- [30] Lei Ma, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Felix Juefei-Xu, Chao Xie, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang. [n. d.]. DeepMutation: Mutation Testing of Deep Learning Systems. In *29th IEEE International Symposium on Software Reliability Engineering (ISSRE), Memphis, USA, Oct. 15–18, 2018*. 100–111.
- [31] Shiqing Ma, Yingqi Liu, Wen-Chuan Lee, Xiangyu Zhang, and Ananth Grama. 2018. MODE: automated neural network model debugging via state differential analysis and input selection. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, 175–186.
- [32] Wei Ma, Mike Papadakis, Anestis Tsakmalis, Maxime Cordy, and Yves Le Traon. 2019. Test Selection for Deep Learning Systems. *arXiv preprint arXiv:1904.13195* (2019).
- [33] JC MacKay David. 1992. A practical bayesian framework for backpropagation networks. *Neural computation* (1992).
- [34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
- [35] Brad L Miller, David E Goldberg, et al. 1995. Genetic algorithms, tournament selection, and the effects of noise. *Complex systems* 9, 3 (1995), 193–212.
- [36] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, and et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.
- [37] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2574–2582.
- [38] Augustus Odena, Catherine Olsson, David Andersen, and Ian Goodfellow. 2019. TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing. In *International Conference on Machine Learning*. 4901–4911.
- [39] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 372–387.
- [40] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*. 582–597.
- [41] Kexin Pei, Yinshi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *SOSP*. 1–18.
- [42] Hung Viet Pham, Thibaud Lutellier, Weizhen Qi, and Lin Tan. 2019. CRADLE: Cross-backend Validation to Detect and Localize Bugs in Deep Learning Libraries. In *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*. IEEE Press, Piscataway, NJ, USA, 1027–1038. <https://doi.org/10.1109/ICSE.2019.00107>
- [43] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8571–8580.
- [44] Jonas Rauber, Wieland Brendel, and Matthias Bethge. 2017. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131* (2017).
- [45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- [46] Lewis Smith and Yarin Gal. 2018. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533* (2018).
- [47] Youcheng Sun, Xiaowei Huang, and Daniel Kroening. 2018. Testing Deep Neural Networks. *arXiv preprint arXiv:1803.04792* (2018).
- [48] Youcheng Sun, Min Wu, Wenjie Ruan, Xiaowei Huang, Marta Kwiatkowska, and Daniel Kroening. 2018. Concolic Testing for Deep Neural Networks. (2018). <https://doi.org/arXiv:1805.00089v2> arXiv:1805.00089

- [49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
- [50] Guan hong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. 2018. Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples. In *Advances in Neural Information Processing Systems 31*. 7728–7739.
- [51] The BBC. 2016. AI image recognition fooled by single pixel change. (2016). <https://www.bbc.com/news/technology-41845878>
- [52] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2018. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *ICSE*. ACM, 303–314.
- [53] Jingyi Wang, Guoliang Dong, Jun Sun, Xinyu Wang, and Peixin Zhang. 2018. Adversarial Sample Detection for Deep Neural Network through Model Mutation Testing. *arXiv preprint arXiv:1812.05793* (2018).
- [54] Model Uncertainty Project Website. 2019. (2019). <https://sites.google.com/view/modeluncertainty>
- [55] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Minhui Xue, Hongxu Chen, Yang Liu, Jianjun Zhao, Bo Li, Jianxiong Yin, and Simon See. 2019. DeepHunter: A Coverage-guided Fuzz Testing Framework for Deep Neural Networks. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2019)*. ACM, New York, NY, USA, 146–157. <https://doi.org/10.1145/3293882>. 3330579
- [56] Xiaofei Xie, Lei Ma, Haijun Wang, Yuekang Li, Yang Liu, and Xiaohong Li. 2019. Diffchaser: Detecting disagreements for deep neural networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 5772–5778.
- [57] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).
- [58] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid. 2018. DeepRoad: GAN-based Metamorphic Testing and Input Validation Framework for Autonomous Driving Systems. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE 2018)*. ACM, New York, NY, USA, 132–142. <https://doi.org/10.1145/3238147.3238187>
- [59] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1, Article 5 (Feb. 2019), 38 pages.
- [60] Yuhao Zhang, Yifan Chen, Shing-Chi Cheung, Yingfei Xiong, and Lu Zhang. 2018. An Empirical Study on TensorFlow Program Bugs. In *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2018)*. 129–140.