

DeepHunter: Hunting Deep Neural Network Defects via Coverage-Guided Fuzzing

Xiaofei Xie¹, Lei Ma², Felix Juefei-Xu³, Hongxu Chen¹, Minhui Xue¹,
Bo Li⁴, Yang Liu¹, Jianjun Zhao⁵, Jianxiong Yin⁶, and Simon See⁶

¹ Nanyang Technological University

² Harbin Institute of Technology

³ Carnegie Mellon University

⁴ University of Illinois at Urbana–Champaign

⁵ Kyushu University

⁶ NVIDIA AI Technology Center

Abstract. In company with the data explosion over the past decade, deep neural network (DNN) based software has experienced unprecedented leap and is becoming the key driving force of many novel industrial applications, including many safety-critical scenarios such as autonomous driving. Despite great success achieved in various human intelligence tasks, similar to traditional software, DNNs could also exhibit incorrect behaviors caused by hidden defects causing severe accidents and losses. In this paper, we propose *DeepHunter*, an automated fuzz testing framework for hunting potential defects of general-purpose DNNs. *DeepHunter* performs metamorphic mutation to generate new semantically preserved tests, and leverages multiple pluggable coverage criteria as feedback to guide the test generation from different perspectives. To be scalable towards practical-sized DNNs, *DeepHunter* maintains multiple tests in a batch, and prioritizes the tests selection based on active feedback. The effectiveness of *DeepHunter* is extensively investigated on 3 popular datasets (MNIST, CIFAR-10, ImageNet) and 7 DNNs with diverse complexities, under large set of 6 coverage criteria as feedback. The large-scale experiments demonstrate that *DeepHunter* can (1) significantly boost the coverage with guidance; (2) generate useful tests to detect erroneous behaviors and facilitate the DNN model quality evaluation; (3) accurately capture potential defects during DNN quantization for platform migration.

Keywords: Deep Neural Networks · Fuzzing · Software Quality Assurance · Coverage Criteria

1 Introduction

Recently great success has been achieved by artificial intelligence (AI) systems, such as IBM’s Watson [1], Amazon Alexa [2], as well as DeepMind’s Atari [3] and AlphaGo [4]. We are now engaged in new AI development and deployment at an unprecedented speed and scale. Deep learning (DL), or deep neural network (DNN) systems, is becoming the paramount ingredient of various kinds of AI-enabled applications, such as speech processing [5], medical diagnostics [6], image processing [7],

and robotics [8], across various implementation platforms such as TensorFlow [9], Keras [10], PyTorch [11]. While DNNs are permeating all industry verticals, it has brought to our attention that DNNs as software 2.0 [12] should be more extensively tested or verified. DNNs definitely deserve more scrutiny than the current practice before they are deployed to safety- and mission-critical applications.

The quality assurance of DNN-based software is still immature, it has caused great losses, such as a Google car accident [13] and an Uber car crash [14]. Systemic and effective testing frameworks for reliably detecting defects and vulnerabilities in real-world sized DNN-based software are in great demand.

In traditional software realm, coverage-guided fuzz (CGF) testing is a well-established technique for defects and vulnerability detection. It helps detect thousands of bugs and vulnerabilities issues in modern software, many of which have been existing for decades [15–21]. CGF performs systematic random mutations on inputs and generates test inputs to drive the software into diverse corner-case states. The major components of the state-of-the-art CGFs often include mutation, feedback guidance, and fuzzing strategy, among which the feedback guidance can provide valuable adjustment to the fuzzing strategy and can significantly improve the efficiency of a fuzzing algorithm.

However, due to the fundamental difference between traditional and DNN based software, traditional fuzzing elements could not be directly applied to DNN fuzzing. For example, the mutations and the feedback are all different in many ways. For traditional software, the mutation is usually rather random and frequently generates *invalid* (or meaningless) seeds that will be rejected by the sanity check in the program quite early. As a result, general-purpose fuzzers usually can only find shallow bugs, such as parsing errors and improper input validations. On the other hand, the input of DNN software typically requires special formats, and inputs that violate the format specifications will be rejected even before the learning procedure starts. Therefore, it is considered more cost-effective to customize a DNN-aware mutation strategy based on some intermediate representations rather than the raw data.

Another challenge in DNN software is that the goal is no longer detecting the *vulnerabilities* or *crashes* in software; rather, we now shift our focus to the *functionality* of DNN results. This is more like a differential testing scenario where we need to additionally distinguish anomaly results from acceptable differences. Furthermore, the study on DNN based software testing is still at its early stage, and whether existing techniques, in particular, *coverage criteria* [22–24], can provide meaningful guidance to DNN fuzzing still lacks extensive and in-depth investigation.

In this work, we are poised to answer these questions and highlight the following contributions:

- We propose a general-purpose coverage guided fuzzing framework *DeepHunter* to systematically test DNN based software, which is among the earliest studies to perform feedback-guided testing for DNNs. The design of *DeepHunter* takes into consideration the unique characteristics of testing DNNs and the scalability towards practical-sized DNNs. (1) In particular, the test execution could be easily paralleled. We propose a batch-based strategy and leverage it to maintain high throughput in obtaining fuzzing results. (2) To enable large-scale automated generation of new test inputs within valid domain, we propose a metamorphic mutation based test genera-

tion technique, which preserves the input semantics before and after mutation. (3) We propose to guide the test generation with pluggable feedback analysis components, including a set of 6 testing criteria of different granularity, to further guide the fuzzing procedures.

- We have performed a large-scale empirical study to evaluate the usefulness of *DeepHunter* in systematically generating tests for coverage enhancement, guided by the 6 recently-proposed coverage criteria.
- We further investigate how each of the very recently-proposed testing coverage criteria helps to guide the fuzzing for, (1) DNN model quality evaluation, (2) error-behavior detection, and (3) defect introduction of quantization for platform migration.

Overall, we find that *DeepHunter* can effectively generate useful tests in general, in terms of (1) improving target coverage, (2) evaluating DNN model quality; (3) detecting erroneous behaviours, as well as (4) capturing the sensitive cases where quantized DNN version fails. To the best of our knowledge, this work is by far among the largest scaled empirical evaluations for DNN testing, using 3 datasets (including ImageNet), 7 DNN models (with large ones like VGG-16, ResNet-50), and a set of 6 coverage criteria as CGF guidance. We will make *DeepHunter* publicly available as an open framework to facilitate further comparative studies on DNN testing.

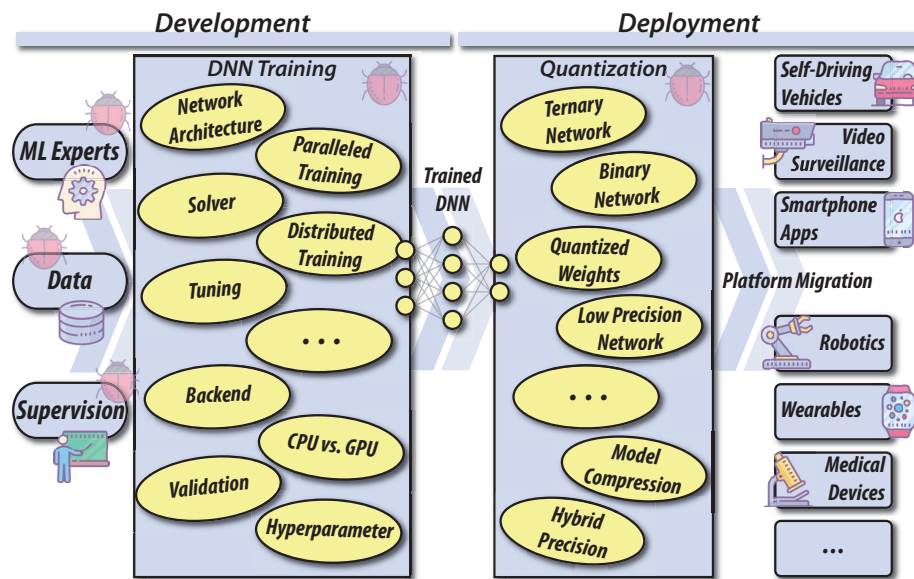


Fig. 1: The development and deployment process for general DNN based software. Our proposed *DeepHunter* is dedicated to assessing the DNN software quality and hunting the defects therein.

2 Preliminaries

2.1 DNN Software Development and Deployment

Over the past several decades, software development methodology [25, 26] has been well-established for traditional software, with many experiences and practices widely applied in software industry. A common software life cycle often consists of several key stages such as requirement analysis, design, implementation, testing, deployment, and maintenance. These development methodology and principles also generally apply to DNN based software. However, different from traditional software, deep learning defines a new data-driven programming paradigm, and keeps its artifact in form of an encoded deep neural network structure and neuron connection weight matrix. Such unique features bring some new challenges for quality assurance of DNN based software, especially in DNN development and deployment phases [27].

Figure 1 gives an overview of the state-of-the-practice DNN software development and deployment. The development phase transforms knowledge of the machine learning experts, the prepared data, and the associated supervision signals into a deep neural network for particular tasks at hand. The training of a DNN involves many tuning knobs, such as the choices of network architecture, backend training framework, solver, and hyper-parameters. Also, one may need to consider the communication overhead, the model parallelization, the data parallelization, the CPU and GPU hybrid training, *etc.*

Once an applicable DNN model is ready, it will oftentimes go through either quantization, or platform migration, or both, before being deployed to end-user applications, such as self-driving cars, video surveillance, smartphones, and wearable devices. This is because the training phase requires a vast amount of computation and energy resources. As the model size and the complexity of the tasks grow, more data are needed to train the network till reaching optimality, which could spend days, if not weeks, in training on high-performance GPU clusters. On the other hand, the deployment of the DNN models is usually into a resource constrained environment with limited computation, storage, and power. Therefore, when migrating from one platform to another, *e.g.*, GPU-cluster trained DNNs to be deployed onto embedded systems or mobile devices, the DNNs usually need to go through a “slimming” process via quantization.

Quantization of DNNs has been widely studied and is considered as one of the most effective approaches to meet the extreme memory and computation requirements that DNNs demand. Studies have shown that to maintain similar level of accuracy and DNN performance, full precision 32-bit floating point weights may not be necessary [28–39]. One can quantize the weights to much lower bits (*e.g.*, from 32-bit floating to 16-bit or to mixed 32 and 16 precision) in order to greatly reduce the model footprint and energy consumption, which has been commonly adopted for industrial level DNN software deployment [40].

However, as depicted in Figure 1, defects might be introduced in both the development and deployment phases. For example, data collection, training program implementation, training execution, *etc.*, could all introduce potential defects at the DNN development stage. Similarly, quantization and platform migration can also introduce defects either due to quantization operator or compatibility issues.

Together, they are among the prime suspects for causing unexpected behaviors and vulnerabilities in DNN software products. The current de facto practice mainly relies on test accuracy to assess the quality of DNNs. However, this is still insufficient especially when the quality of test data is low. A low quality test only *partially* measures the DNN quality, and is unsuitable to provide insights to defects and vulnerabilities DNN software, causing some fatal defects missed without giving any feedback to the DNN developers.

2.2 Coverage-based Grey-box Fuzzing

Fuzzing has gained its popularity in academia and industry due to its scalability and effectiveness in generating useful tests for defect detection. Based on awareness of the target program structure, fuzzers can be classified as black-box [41], white-box [42] or grey-box [15]. One of the most successful techniques is coverage-based grey-box fuzzing (CGF), which strikes a balance between effectiveness and efficiency by using code coverage as feedback. Many state-of-the-art CGFs, such as AFL [15], libFuzzer [16] and VUzzer [19], have been widely used and proven to be effective.

Given a target program, CGF uses a lightweight instrumentation to collect the coverage information during fuzzing. A typical CGF usually performs the following loop [43]: (1) selecting seeds from the seed pool; (2) mutating the seed a certain number of times to generate new tests with mutation strategies such as bitwise/bytewise flips, block replacement, and crossover on two seed files; (3) running the target program against the newly generated inputs, and recording the executed traces; (4) reporting fault seeds if crashes are detected, and saving those interesting seeds that cover new traces into the seed pool. Such iteration continues until given computation resource exhausts. The two key components in CGF are *mutation* and *coverage feedback* that largely determine the efficiency of fuzzing.

Despite the huge differences between traditional programs and DNNs, the success of CGF on the former still gains insight into the fuzzing on the latter. For example, the target traditional program mirrors the DNN, the seed of fuzzer mirrors the input of the DNN, and the coverage feedback could be some coverage of DNN. Considering the unique characteristics of the DNN, it is still challenging to develop effective *mutation strategies* and *coverage criteria* in terms of DNN fuzzing. This paper aims to fill this gap by designing effective CGF framework towards providing a quality assurance gadget during the DNN development and deployment process.

3 METHODOLOGY

In this section, we elaborate proposed coverage guided fuzzing for DNNs. We take an overview of *DeepHunter* and then describe each of the key components in details.

3.1 Overview of DeepHunter

Fig. 2 depicts the overview of *DeepHunter*, and Algorithm 1 specifies the details. At a high level, *DeepHunter* consists of three major components: *Metamorphic Mutation*,

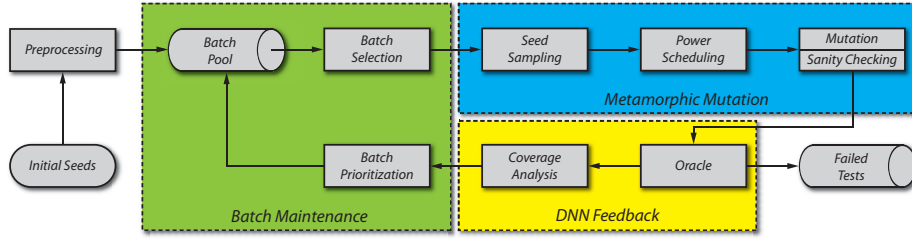


Fig. 2: The workflow of *DeepHunter*, which leverages metamorphic mutation to generate tests with coverage feedback as guidance.

DNN Feedback, and *Batch Pool Maintenance*. We define an atomic input of the DNN (e.g., an image) as a *seed* and a set of seeds (e.g., multiple images) as a *batch*. As DNNs can quickly predict multiple seeds (i.e., a batch) at once, we maintain a batch pool instead of a seed pool to improve the fuzzing effectiveness. During fuzzing, *DeepHunter* first selects a batch and generates a large number of mutated seeds, then the DNN predicts all mutated seeds *at once*. At last, *DeepHunter* maintains the pool based on the coverage information. The workflow of *DeepHunter* is detailed below (see Algorithm 1).

The inputs of *DeepHunter* are the initial seeds and the target DNN model under test. Before the fuzzing loop, initial seeds are constructed as batches, which are added into the batch pool (Line 2). During fuzzing process, the fuzzer selects one batch from the priority batch pool (Line 3). From the batch, the fuzzer samples some seeds to be mutated (Line 4). The fuzzer applies a power scheduling against the sampled seeds to determine the mutation chances for each seed (c.f. Section 3.3). For each sampled seed, the fuzzer will mutate it for the assigned times (Line 8), and sanitize each mutated seed (c.f. Section 3.2) since random mutation may generate some meaningless seeds (e.g., images imperceptible to the human eye). For each valid mutant of the original seeds, the test oracle will verify whether this is a failed test. After mutating all sampled seeds, the survived mutants are constructed as a batch. DNNs will predict all the seeds and collect the coverage information of the batch (Line 14). If the batch gains the coverage, it will be added into the batch pool. Batch prioritization will prioritize the batches that have been seldomly fuzzed (Lines 16-17, c.f. Section 3.4 and Section 3.5).

3.2 Transformation and Mutation

Traditional fuzzers such as AFL mutate the input with bitwise / bytewise flips, block replacement, crossover between input files, etc. However, these strategies usually generate too many inputs that are meaningless in DNN fuzzing. For example, images or voices that are imperceptible to human senses should be discarded from the mutation. Hence, one challenge is how to balance between increasing the changeability of mutation and generating meaningful inputs. If the mutation change is very small, the newly generated input may be almost unchanged; despite the fact that it may be meaningful, the fuzzer has lower chances of finding failed tests. On the other hand, if the mutation change is very large, more failed tests may be identified; however, the failed tests are more likely to be meaningless.

Algorithm 1: DeepHunter

input : I : Initial Seeds, DNN , Target Neural Network
output: F : Failed Tests
const : K : Total number of mutation for a batch

```
1  $F \leftarrow \emptyset$ ;  
2  $T \leftarrow Preprocess(I)$ ;  
3 while  $B \leftarrow SelectNext(T)$  do  
4    $S \leftarrow Sample(B)$ ;  
5    $P_S \leftarrow PowerSchedule(S, K)$ ;  
6    $B' \leftarrow \emptyset$ ;  
7   for  $\mathcal{I} \in S$  do  
8     for  $i$  from 1 to  $P_S(\mathcal{I})$  do  
9        $\mathcal{I}' \leftarrow Mutate(\mathcal{I}, B)$ ;  
10      if  $isFailedTest(\mathcal{I}')$  then  
11         $F \leftarrow F \cup \{\mathcal{I}'\}$ ;  
12      else if  $isChanged(\mathcal{I}, \mathcal{I}')$  then  
13         $B' \leftarrow B' \cup \{\mathcal{I}'\}$ ;  
14    $cov \leftarrow Predict(DNN, B')$ ;  
15   if  $CoverageGain(cov)$  then  
16      $T \leftarrow T \cup \{B'\}$ ;  
17      $BatchPrioritize(T)$ ;
```

In this work, we mainly focus on image inputs. To solve the aforementioned challenge, we develop a metamorphic mutation strategy. The basic objective is that *given an image i , the mutator generates another new image i' such that the semantics of i and i' are the same from the perspective of people.*

Image Transformation. To increase the changeability of mutation, we select eight image transformations which are classified into two categories:

- *Pixel Value transformation* \mathcal{P} : change image contrast, image brightness, image blur and image noise.
- *Affine transformation* \mathcal{G} : image translation, image scaling, image shearing and image rotation.

Intuitively, *Pixel Value transformation* changes the pixel values of the image while *Affine transformation* moves the pixels of the image. The transformations have been proved to be effective and useful in [23].

Definition 1. An image \mathcal{I}' is one-time mutated from \mathcal{I} if \mathcal{I}' is generated after a transformation t on \mathcal{I} (denoted as $\mathcal{I} \xrightarrow{t} \mathcal{I}'$), where $t \in \mathcal{P} \cup \mathcal{G}$. An image \mathcal{I}' is sequentially mutated from \mathcal{I} if \mathcal{I}' is generated after a sequence of one-time mutations ($\mathcal{I} \xrightarrow{t_0} \mathcal{I}_1, \mathcal{I}_1 \xrightarrow{t_1} \mathcal{I}_2, \dots, \mathcal{I}_n \xrightarrow{t_n} \mathcal{I}'$) (denoted as $\mathcal{I} \xrightarrow{t_0, t_1, \dots, t_n} \mathcal{I}'$).

Metamorphic Mutation. By setting proper parameters for different transformations, it is assumed that the image after *one-time* mutation has the same semantics with the

Algorithm 2: Mutate

```
input :  $\mathcal{I}$ : Seed
output:  $\mathcal{I}'$ : New Seed
const :  $TRY\_NUM$ : The maximum number of trials
1  $(\mathcal{I}_0, \mathcal{I}'_0, state) \leftarrow info(\mathcal{I})$ ;
2 for  $i$  from 1 to  $TRY\_NUM$  do
3   if  $state == 0$  then
4      $t \leftarrow randomPick(\mathcal{G} \cup \mathcal{P})$ ;
5   else
6      $t \leftarrow randomPick(\mathcal{P})$ ;
7    $p \leftarrow pickRandomParam(t)$ ;
8    $\mathcal{I}' \leftarrow t(\mathcal{I}, p)$ ;
9   if  $isSatisfied(f(\mathcal{I}'_0, \mathcal{I}'))$  then
10    if  $t \in \mathcal{G}$  then
11       $state \leftarrow 1$ ;
12       $\mathcal{I}'_0 \leftarrow t(\mathcal{I}_0, p)$ ;
13       $info(\mathcal{I}') \leftarrow (\mathcal{I}_0, \mathcal{I}'_0, state)$ ;
14      return  $\mathcal{I}'$ ;
15 return  $\mathcal{I}$ ;
```

original image. However, during fuzzing, one image can be sequentially mutated from the original image, it is challenging to generate meaningful images after a sequence of mutations. To boost the mutation effectiveness, we propose the metamorphic mutation.

In order to ensure the meaningfulness of the mutated image as much as possible, we adopt a conservative strategy that makes *Affine Transformation* to be selected only once because multiple affine transformations are more likely to generate meaningless images. We assume that an affine transformation will not affect the semantics under the selected parameters. *Pixel Value Transformation* can be selected multiple times and we use L_0 and L_∞ to limit the pixel-level change. Suppose an image \mathcal{I} is mutated to \mathcal{I}' by a pixel value transformation, then \mathcal{I}' is meaningful in terms of \mathcal{I} if $f(\mathcal{I}, \mathcal{I}')$ (Equation 1) is satisfied.

$$f(\mathcal{I}, \mathcal{I}') = \begin{cases} L_\infty \leq 255, & \text{if } L_0 < \alpha \times \text{size}(\mathcal{I}) \\ L_\infty < \beta \times 255, & \text{otherwise} \end{cases} \quad (1)$$

where $0 < \alpha, \beta < 1$, L_0 represents the maximum number of the changed pixels, L_∞ represents the maximum value of the pixel changes, $\text{size}(\mathcal{I})$ is the number of pixels in image $0 < \mathcal{I}$.

Intuitively, if the number of changed pixels is very small ($< \alpha \times \text{size}(\mathcal{I})$), we assume it does not change the semantics and L_∞ can be any value. If the number of changed pixels exceeds the boundary, we limit the maximum changed value ($< \beta \times 255$).

Definition 2. Given a mutated image \mathcal{I} , the original image (denoted as \mathcal{I}_0) of \mathcal{I} is the image in the initial seeds and \mathcal{I} is one-time mutated or sequence mutated from \mathcal{I}_0 , i.e., $\mathcal{I}_0 \xrightarrow{t_0, \dots, t_n} \mathcal{I}$, where $n \geq 0$. The reference image (denoted as \mathcal{I}'_0) is defined as:

$$\mathcal{I}'_0 = \begin{cases} \mathcal{I}_j, \exists 0 \leq j \leq n. t_j \in \mathcal{G} \wedge \mathcal{I}_0 \xrightarrow{t_0, \dots, t_j} \mathcal{I}_j \\ \mathcal{I}_0, \text{ otherwise} \end{cases}$$

Algorithm 2 shows the details of the mutation, which takes an original image \mathcal{I} as the input and the mutated image \mathcal{I}' as the output. We first obtain the tuple $(\mathcal{I}_0, \mathcal{I}'_0, state)$ (Line 1) which is recorded in the batch. *state* is the current mutation state 0 or 1, which represents whether an *Affine Transformation* is used. *DeepHunter* tries to mutate a meaningful image \mathcal{I}' with a maximum number of trials *TRY_NUM* (Line 2-14). It randomly picks a transformation t . If the current mutation state is 0, it can select from both *Affine Transformation* and *Pixel Value Transformation* (Line 4). If the mutation state is 1, it can only use a pixel value transformation (Line 6). For the transformation t , it picks a parameter randomly (Line 7) and performs the transformation (Line 8).

Next, Algorithm 2 computes L_0 and L_∞ between the reference image \mathcal{I}'_0 and the new mutated image \mathcal{I}' to check whether \mathcal{I}' is meaningful (Line 9). Note that we compare \mathcal{I}' with reference image \mathcal{I}'_0 instead of original image \mathcal{I}_0 because: (1) the pixels between \mathcal{I}'_0 and \mathcal{I}' are corresponding, which is necessary to compute L_0 and L_∞ and (2) we assume that \mathcal{I}'_0 and \mathcal{I}_0 have the same semantics under our conservative parameters. Hence, if $f(\mathcal{I}'_0, \mathcal{I}')$ (c.f. Equation 1) is satisfied, we can conclude that \mathcal{I}' and \mathcal{I}_0 also have the same semantics and the mutation is successful. If the selected t is an *Affine Transformation*, it updates the mutation state of \mathcal{I}' and \mathcal{I}'_0 (Line 11-12). At last, Algorithm 2 saves the current image \mathcal{I}' (Line 13) and ends the mutation (Line 14). If there is no successful mutation after *TRY_NUM* mutations, it outputs the image \mathcal{I} .

3.3 Power Scheduling

As described in Algorithm 2, *DeepHunter* mutates one image with a limited number of tries. If $f(\mathcal{I}'_0, \mathcal{I}')$ is satisfied, the mutation of \mathcal{I} is successful. Actually, the possibility of successful mutation depends on the difficulty that $f(\mathcal{I}'_0, \mathcal{I}')$ is satisfied.

Given an image \mathcal{I} and its reference image \mathcal{I}'_0 , we define its *mutation potential* as $\beta \times 255 \times \text{size}(\mathcal{I}) - \text{sum}(\text{abs}(\mathcal{I} - \mathcal{I}'_0))$. Intuitively, the mutation potential approximately represents the mutation space of an image \mathcal{I} , i.e., the difficulty that $f(\mathcal{I}'_0, \mathcal{I}')$ is satisfied. $\beta \times 255 \times \text{size}(\mathcal{I})$ represents the maximum value that the image can change. $\text{sum}(\text{abs}(\mathcal{I} - \mathcal{I}'_0))$ represents the value that the image has changed in terms of \mathcal{I}'_0 . For example, suppose \mathcal{I}' is sequentially mutated from $\mathcal{I}'_0 : (\mathcal{I}'_0 \xrightarrow{t_0} \mathcal{I}_1, \dots, \mathcal{I}_n \xrightarrow{t_n} \mathcal{I}')$, the mutation potential of images in the front of the sequence (e.g., \mathcal{I}_1) is more likely to be higher than those in the tail (e.g., \mathcal{I}').

The power scheduling is a procedure for *DeepHunter* to decide mutation chances for different seeds (i.e., images). To boost the efficiency of fuzzing, we expect to mutate more images that have higher mutation potential.

Table 1: The plugable coverage criteria integrated in *DeepHunter* for test guidance. Besides the first five criteria originally proposed in [22, 24], we also include an extra BKNC criterion that plays as a counterpart of TKNC by measuring the ratio of top-k most hypoactivated neurons.

Subject Cov. Criteria	Description
Neuron Cov. (NC)	The ratio of activated neurons
K-multisec. Neu. Cov. (KMNC)	The ratio of covered k-multisections of neurons
Neuron Bound. Cov. (NBC)	The ratio of covered boundary region of neurons
Strong Neuron Act. Cov. (SNAC)	The ratio of covered hyperactive boundary region
Top-k Neu. Cov. (TKNC)	The ratio of neurons in top-k hyperactivated state
Bottom-k Neu. Cov. (BKNC)	The ratio of neurons in top-k hypoactivated

3.4 Plugable Coverage-Guided Fuzzing

A dumb fuzzer without any coverage guidance aimlessly mutates the seed, without knowing whether the generated test input is preferable. Consequently, such a fuzzer may frequently keep seeds that do not bring new desired information; even worse, mutation on these seeds may bury other “interesting” seeds in the fuzzing queue, significantly decreasing the fuzzing effectiveness. Therefore, modern fuzzers for traditional software often embrace some feedback such as *code coverage*.

In this paper, *DeepHunter* selects six different criteria [22, 44] (Table 1) as different feedback to determine whether the newly generated batch should be kept for further mutation. The criteria have been proven to be useful to capture the internal DNN states. However, due to the huge numerical space of each neuron value and the large scale nature of the DNN software, the fuzzer could be overloaded with *flooded feedback*. In fact, without triaging, seed inputs with similar neuron values will be *unnecessarily* retained. Due to the mutation instinct, there will be a huge number of such mutants that originate from a given seed. To tackle this issue, we equally split the numerical neuron-feedback interval of each criteria into different buckets, each of which will be regarded as an “equivalent class”. If a new seed with its coverage results of a neuron falling into existing buckets, it is out of the interest and discarded. This mechanism is inspired from the “loop bucket” practice used in the traditional fuzzing framework (*e.g.*, AFL), to mitigate the trace exploitation issue [45].

3.5 Batch Prioritization

Batch prioritization decides which batch should be picked next. We adopt a strategy which probabilistically selects the batch based on the number of times it has been fuzzed. Specially, the probability is computed by:

$$P(B) = \begin{cases} 1 - f(B)/\gamma, & \text{if } f(B) < (1 - p_{\min}) \times \gamma \\ p_{\min}, & \text{otherwise} \end{cases} \quad (2)$$

where B is a batch, $f(B)$ represents how many times the batch B has been fuzzed and $p_{\min} > 0$ is the minimum probability. The values of parameters γ and p_{\min} can be adjusted.

Table 2: Subject datasets and DNN models.

DataSet	Dataset Description	DNN Model	#Neuron	#Layer	Test Acc.
MNIST	Hand written	LeNet-1	52	7	0.976
	digits recog.	LeNet-4	148	8	0.989
	from 0 to 9	LeNet-5	268	9	0.990
CIFAR-10	General image	ResNet-20	2,570	70	0.917
	with 10-class	VGG-16	12,426	17	0.928
ImageNet	1000-class large	MobileNet	38,904	87	0.871*
	scale image cla.	ResNet-50	94,059	176	0.929*

* The reported top-5 test accuracy of pretrained DNN model in [46].

The basic idea here is to prioritize the batches that have been seldomly fuzzed. For example, the probability of new mutated batch is 1 since it gains new coverage and is regarded as interesting. To keep the diversity, other batches that have been fuzzed many times also have a minimum probability p_{\min} to be selected.

4 Experiments

DeepHunter is implemented in Python and C: the metamorphic mutation component and DNN coverage feedback component are implemented in Python based on deep learning framework Keras (ver.2.1.3) [10] with TensorFlow (ver.1.5.0) backend [9]; the batch maintenance component is implemented in C for efficiency. We evaluate *DeepHunter* by investigating the following research questions:

RQ 1: What coverage can *DeepHunter* achieve when guided by the six testing criteria?

RQ 2: Does *DeepHunter* facilitate the DNN model evaluation effectively?

RQ 3: Can *DeepHunter* enable diverse erroneous behavior detection of DNNs?

RQ 4: Can *DeepHunter* detect potential defects introduced during DNN quantization?

4.1 Datasets and DNN Models

We select three popular publicly available datasets (*i.e.*, MNIST [47], CIFAR-10 [48], and ImageNet [49]) as the evaluation subject datasets (see Table 2). For each dataset, we study popular DNN models [50–52] that are widely used in previous work. In particular, we perform extensive controlled study on MNIST and CIFAR-10, and investigate the scalability and usefulness of *DeepHunter*. Table 2 summarizes the structures and complexity of the DNNs used in this paper.

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 in 10 classes (*i.e.*, handwritten digits from 0 to 9). Each MNIST image is a single-channel of size $28 \times 28 \times 1$. On MNIST, we have studied three LeNet family models (LeNet-1, LeNet4, LeNet-5 [50])

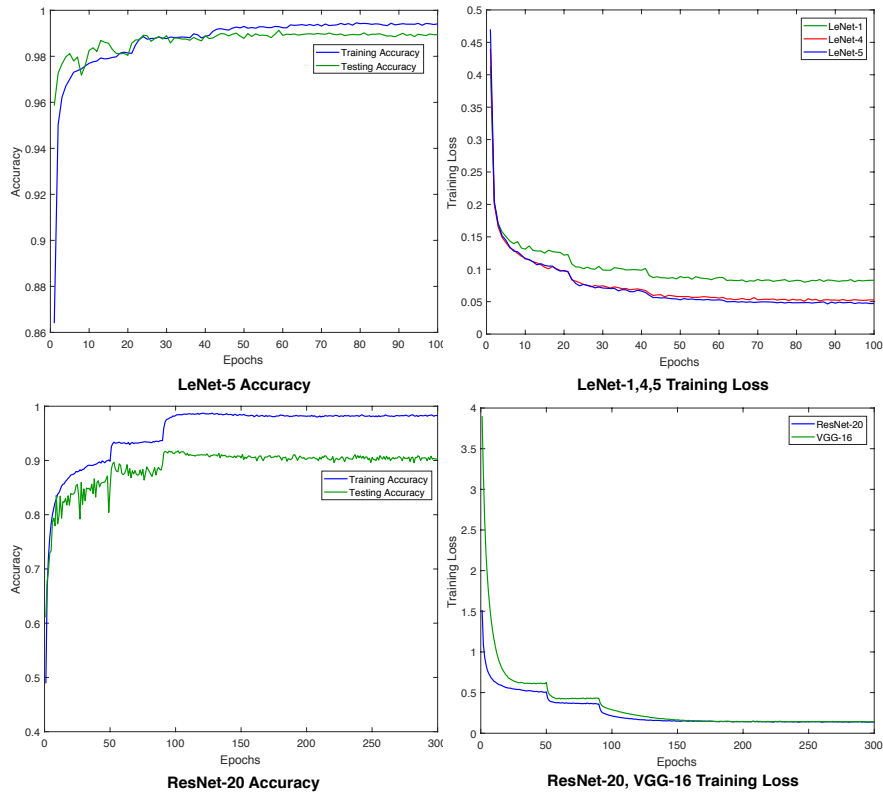


Fig. 3: Controlled results on training loss, training accuracy, and test accuracy.

as the subject models. We train each of the LeNet models in a controlled setting, with detailed configuration and discussion in Section 4.4.

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.*, airplanes, cars, birds, and cats). Each CIFAR-10 image is three-channel of size $32 \times 32 \times 3$. The classification task of CIFAR-10 is generally harder than that of MNIST due to the data size and complexity. To obtain competitive performance on CIFAR-10, we study two well-known DNN models (*i.e.*, ResNet-20 [53] and VGG16 [54]) as the subject models.

ImageNet. To further demonstrate that *DeepHunter* scales to practical-sized dataset and DNN models, we also select ImageNet, which is a large-scale visual recognition challenge (ILSVRC) dataset for general-purpose image classification. The complexity of ImageNet is characterized by a large number of training data (*i.e.*, over one million) and test data (*i.e.*, 50,000), as well as large data points, each of which is of size $224 \times 224 \times 3$ ($\sim 50x$ dimensionality of CIFAR-10). Therefore, it would be an ordeal for any automated testing tool to work on ImageNet sized dataset and DNN models. Specifically, we try to examine whether *DeepHunter* enables the fuzz testing on ImageNet dataset and practical-sized DNN models (*i.e.*, VGG-19 [54], ResNet-50 [53]).

Table 3: DNN model and their training and test performance.

DataSet	DNN	Epoch	Syno.	Train Loss	Train Acc.	Test Acc.
MNIST	LeNet-1	10	A	0.131	0.965	0.967
		30	B	0.099	0.975	0.975
		45	C	0.087	0.979	0.976
	LeNet-4	10	A	0.117	0.974	0.978
		25	B	0.077	0.986	0.986
		50	C	0.058	0.990	0.989
	LeNet-5	10	A	0.116	0.977	0.983
		30	B	0.071	0.988	0.989
		45	C	0.056	0.992	0.990
CIFAR-10	ResNet-20	40	A	0.515	0.894	0.859
		55	B	0.385	0.932	0.880
		95	C	0.239	0.977	0.917
	VGG-16	30	A	0.623	0.914	0.850
		55	B	0.443	0.965	0.900
		95	C	0.316	0.995	0.928

4.2 Experiment Setup

Subject DNN Model Training and Preparation. Since the DNN model quality could affect the evaluation results, we carefully select the well-known DNN models that obtain competitive performance on each studied dataset. In this paper, we closely follow the common machine learning training practice and instructions [55, 56], and set up a three-stage adaptive learning rate for training MNIST and CIFAR-10 DNN models. The larger learning rate at the early training stage accelerates the training convergence, while the later stage with a smaller learning rate allows the performance fine-tuning. The training loss, training accuracy, and test accuracy for each model are shown in Figure 3. From the training accuracy and training loss curve, we can observe that the training process jumps into three different stages as expected. We follow the machine learning practice and select the best candidate models with the most competitive performance without overfitting as the subject DNN model instances for fuzz testing (see Table 1 and DNN C variants in Table 3 with the detailed epochs and training information). Due to the large size of training data and training effort of ImageNet, we select the pretrained MobileNet and ResNet-50 [46] as the subject models [53].

For RQ 2, we try to evaluate whether *DeepHunter* enables the DNN model quality evaluation. Therefore, besides the best candidate C instances for each model used for MNIST and CIFAR-10, we also select other two instances A and B from each of the first two training stages, which allows us to sort the model quality relation $Q_A < Q_B < Q_C$ as the groundtruth.⁷ The selected groundtruth model instances for quality evaluation is summarized in Table 3.

Coverage-Guided Fuzz Testing. After all DNN models are obtained for each dataset (see Table 2), we use *DeepHunter* to perform large-scale fuzz testing on each of these models to generate tests. For each DNN model, we randomly sample tests as initial seed batches from their original test dataset such that all these tests are correctly handled

⁷The general DNN quality groundtruth is hard to obtain; the state of the practice still relies on test accuracy. Our three-stage training procedure follows the machine learning practice [55, 56] and allows to obtain model instances, each from one of the three training stages, so that we could obtain the desired model quality relation with high confidence.

by the model.⁸ Furthermore, on MNIST and CIFAR-10 datasets, the sampled initial seeds are also correctly handled by each of the model instance A and B, as shown in Table 3. This allows us to perform the controlled evaluation (for RQ 2) on the usefulness of *DeepHunter* for quality evaluation of models, where the initial seeds fail to distinguish each of the model instances A, B, C in terms of prediction accuracy. We configure *DeepHunter* to use each of the six studied testing criteria for guided fuzz testing with a time budget of 24 hours, allowing the testing coverage to achieve relatively saturated status. The obtained tests are used to perform post-phase analysis for each research question. Note that the post-phase analysis for all the research questions are also computational intensive. To support such a large-scale subject DNN model set training, which are fuzz testing and post-phase analysis, we run all the experiments on a high performance computer cluster. Each cluster node runs a GNU/Linux system with Linux kernel 4.4.0 on a 28-core 2.3GHz Intel Xeon 64-bit CPU with 196 GB of RAM equipped with a NVIDIA Tesla V100 16G-GPU.

4.3 Coverage Results Guided by Different Testing Criteria

To answer RQ 1, the achieved coverage of *DeepHunter* guided by different testing criteria is shown in Table 4. *DeepHunter* generates tests that significantly increase the corresponding coverage compared with initial seeds, as confirmed by Wilcoxon Signed Ranks Test ($p < 0.01$) for all cases.

The difficulty to cover different criteria is different. For example, the TKNC obtained by initial seeds has already obtained very high coverage in some cases (see LN-5 with 76.3% initial NC), While the initial seeds only obtain 0.1% for ResNet-20 on NBC and SNAC. Such results are consistent with the coverage trends reported in [24], where the NBC and SNAC are more challenging to cover since they represent the corner-regions where neuron states go beyond a normal region. Even though, *DeepHunter* is still able to significantly boost such criteria, increased by 15.86x (from 0.7% to 10.3% on MobileNet) to 77.5x (from 0.04% to 3.1% on VGG-16).

Answer to RQ 1: *DeepHunter* significantly boosts the coverage across with different criteria guidance.

4.4 DNN Model Quality Evaluation

Although software quality standards and metrics are well-established for traditional software, the DNN quality assurance research is still at an early stage, with most current work relying on test accuracy. To answer RQ 2, we investigate whether *DeepHunter* provides more useful feedback and evaluation on the DNN model quality, by using controlled setting on MNIST and CIFAR-10 (see Table 3). Note that the initial seeds are all correctly predicted by all instances A, B, C of each model.

We have used *DeepHunter* to generate tests for instance C of each model. Among these tests, we keep those correctly predicted by C, and run these tests on instances A

⁸We sample 1,000 initial seeding data in 10 batches (each contains 100 test data) for MNIST and CIFAR-10, and 500 seeding data for ImageNet in 20 batches w/ equal size.

Table 4: The coverage of initial seeds and tests with *DeepHunter* guided by the corresponding testing criteria.

DNN Model	NC(%)		KMNC(%)		NBC(%)		SNAC(%)		TKNC(%)		BKNC(%)	
	Init.	D.H.	Init.	D.H.	Init.	D.H.	Init.	D.H.	Init.	D.H.	Init.	35.4
LN-1	22.9	31.3	31.4	92.2	1.2	24.4	0.3	22.1	49.7	49.8	49.8	49.8
LN-4	56.3	61.1	21.9	75.9	0.4	15.1	0.2	20.2	69.5	72.7	27.2	31.1
LN-5	58.0	70.8	20.6	78.4	0.3	11.3	0.2	19.2	76.3	83.2	24.6	30.5
RN-20	7.5	10.8	36.0	75.1	0.1	8.1	0.1	8.11	62.3	68.0	63.3	68.7
VGG16	41.9	46.2	41.1	84.1	0.04	3.1	0.1	3.1	13.3	15.1	13.9	16.2
MN	7.0	7.8	26.9	73.4	0.7	10.3	0.4	9.5	4.9	7.9	5.2	8.5
RN-50	4.8	9.4	23.9	51.6	0.1	3.2	0.1	3.8	14.1	22.9	13.7	20.6
Avg.	28.3	33.9	28.8	75.8	0.4	10.8	0.2	12.3	41.4	45.6	28.2	32.2

Table 5: The controlled DNN model instance quality evaluation accuracy results.

DNN Instan.	NC(%)		KMNC(%)		NBC(%)		SNAC(%)		TKNC(%)		BKNC(%)	
	A	B	A	B	A	B	A	B	A	B	A	B
LN-1	99.0	99.5	92.8	97.5	93.0	96.5	90.1	96.0	91.5	98.5	95.4	98.7
LN-4	98.1	99.6	92.7	97.3	87.1	95.5	91.0	95.4	91.5	95.7	89.3	95.9
LN-5	95.4	97.7	91.3	94.1	88.4	96.9	91.1	97.0	92.4	97.0	91.3	96.7
RN-20	92.6	93.9	81.7	87.2	83.8	87.4	83.8	86.3	81.9	87.1	85.3	87.8
VGG16	86.0	87.7	78.0	83.3	80.0	81.3	82.9	83.9	82.8	83.3	84.1	85.2
Avg.	94.2	95.7	87.3	91.9	86.4	91.5	87.8	91.7	88.0	92.3	89.1	92.9

and B of each model. The obtained accuracy for instances A and B of each model is shown in Table 5. We can see that *DeepHunter* facilitates the DNN model quality evaluation, and the quality evaluation results are largely consistent with the model quality groundtruth (i.e., $Q_B > Q_A$). The tests generated by different coverage guidance exhibit different abilities to show the model quality difference. For example, on LeNet-5, the NC only slightly shows B might have better quality; however, this becomes obvious when it comes to NBC, where instance A achieves 88.4% accuracy and instance B achieves 96.9%.

We see that most of the accuracy of instances A and B under our generated tests (see Table 5) are *lower than* the original test accuracy (see Table 2). On the contrast, most of the absolute accuracy differences between instances A and B under our generated tests *outnumber* those under original test data. This indicates that the generated tests can better distinguish the qualities of instances A and B, and instance C is able to generate high quality tests than the other two, which is consistent with our expectations.

Answer to RQ 2: *DeepHunter* facilitates the model quality evaluation through guided fuzz testing. The tests generated with different coverage guidance exhibit different test capabilities, providing different feedback to the model quality.

4.5 DNN Erroneous Behavior Detection

To answer RQ 3, during the fuzz testing process of *DeepHunter*, we continuously collect the generated tests that trigger erroneous behaviors of DNNs. Since our metamorphic mutation performs constraint-based transformation on inputs, to ensure no changes

Table 6: The number of unique error triggering tests generated by *DeepHunter* with different coverage guidance.

DNN Models	Unique Error Triggering Tests (unit in 1 k)					
	NC	KMNC	NBC	SNAC	TKNC	BKNC
LeNet-1	6.9	1.1	6.7	8	6.8	8.6
LeNet-4	4.7	0.6	2.9	3.3	2.2	4.5
LeNet-5	2.9	0.7	3.1	3.3	1.3	3.2
RN-20	0	7.0	7.8	7.9	6.1	7.2
VGG-16	2.2	6.3	8.1	8.5	6.8	8.3
MobileNet	1.5	10.8	11.6	9.6	16.6	13.8
RN-50	1.3	8.1	8.8	8.8	9.7	8.6
SUM	19.5	34.6	49	49.4	49.5	54.2

of the semantic between the original image and the transformed one, we perform prediction check on images before and after transformation in batch and record the tests that trigger the erroneous behaviors of the DNN under test. The detected erroneous behaviors from proposed coverage criteria for each model is shown in Table 6.⁹ Consequently, it is not surprising that *DeepHunter* successfully generates tests to trigger the erroneous behaviors of DNNs. The recent work [22, 23] have already shown that testing only based on neuron coverage already generates thousands of erroneous triggering tests.

There appears a case on RN-20 with the neuron coverage 0. After generating 24 batches, the tests generation converges (*i.e.*, new seeds cannot cover new coverage in terms of NC). Thus the metamorphic mutation is always run on existing batches, and new seeds are always *one-time* mutated from existing batches. As a result, in this case, one transformation under our conservative strategy (*c.f.* Section 3.2) is difficult to generate erroneous triggering tests.

Answer to RQ 3: *DeepHunter* can effectively generate tests to trigger erroneous behaviors of the DNN under tests, which also scales well to practical-sized datasets and DNN models.

4.6 Defect Detection under Controlled DNN Quantization Settings

To answer RQ 4, recently there exists a strong demand to deploy DNN solutions on diverse platforms such as mobile device, edge computing device. Due to the computation and power limitation, a common practice is to quantize the DNN model from high precision floating to a lower precision form, to reduce the size for deployment. However, the quantization could introduce potential unexpected erroneous behaviors. An effective test suite should be able to capture such error cases as feedback to DL developer for further analysis and debugging. In this research question, we investigate whether *DeepHunter* is useful to detect potential defects during quantization.

For each of studied DNN model in Table 2 (that is 32-bit floating point precision), we perform quantization with 3 configurations: (1) randomly sample 1% of weights

⁹Note that once error-trigger tests are generated, they are recorded for further processing without putting them back into the batch pool.

to truncate 32-bit floating point to 16-bit, resulting a mixed precision DNN model, (2) randomly sample 50% weights to truncate 32-bit floating point to 16-bit, and (3) truncate all weights from 32-bit floating point to 16-bit.¹⁰

Notice that the initial seeds of each dataset cannot detect the erroneous behavior before and after quantization. Then, we reuse the tests generated by DeepHunter to evaluate quantized models, the results are summarized in Table 7. In all cases, *DeepHunter* enables to detect the potential minor erroneous behaviours introduced during quantization.

In many of the configurations, the tests generated with NBC and SNAC guidance detect more erroneous issues. One potential reason is that the tests with higher NBC and SNAC tends to cover the corner-region behavior of neurons, which could potentially trigger the erroneous behaviors of quantized model. Another interesting finding we found that the number error trigger tests for full quantization DNN model could sometimes be smaller than the mix-precision quantization counterparts. For example on LeNet-4 TKCN configuration, we found that 31 erroneous behavior on full quantization DNN model, while averaged 33 erroneous behavior on 50%. Intuitively, the large quantization ratio, more weights lose precision, and more erroneous behavior could be introduced. However, our evaluation results hints that sometimes the error introduced by more weight precision loss might cancel each other and obtain an less erroneous quantized version.

Answer to RQ 4: *DeepHunter* can effectively detect potential defects introduced during DNN quantization, albeit a minor precision loss.

4.7 Discussion and Threats to Validity

We perform extensive study on fuzz testing using 6 coverage criteria for guidance. In this section, we tend to discuss the potential effects of studied coverage criteria as feedback to guide fuzz testing, based on our experimental results.

From the coverage results (*c.f.* Table 4) as well as the corresponding criteria definition, we find that KMNC is a fine-grained criterion, representing k -multisection of neurons, which easily facilitates to generate interesting tests. For example, the average coverage gain of KMNC is 47%, outnumbering the others. On the other hand, NC is a relatively coarse-grained criterion which represents records the ratio of activated neurons. Due to this (see Table 5), the fuzz testing guided by NC cannot generate effective results to evaluate the models with various quality. The results in Table 2 and Table 3 also show that NC is less effective in error triggering test detection and sensitive defect detection.

In comparison to KMNC and NC, the other four criteria show different behavior to guide fuzz testing, some of which could be difficult to cover such as SNAC and NBC. They tend to guide fuzz testing in generating corner-case tests, so that to trigger more erroneous behaviors. In Table 2, more error-triggering tests are generated by the four than those by KMNC and NC in many cases. Meanwhile, KMNC is a fine-grained coverage

¹⁰Due to randomness of the first two configurations, we repeat the sampling procedure 5 times to average the results.

Table 7: The number of sensitive defects are detected by *DeepHunter* during DNN model quantization, with full quantization from 32-bit to 16-bit floating conversion, as well as with mixed precision with random parts of weights quantized. The number of defects for 1% and 50% quantization ratio are averaged detected defects over five runs.

DNN Models	Quan. Ratio (%)	Number Defects Detected					
		NC	KMNC	NBC	SNAC	TKNC	BKNC
LeNet-1	1	9	17	61	63	18	21
	50	43	79	111	141	75	67
	100	36	77	107	164	82	62
LeNet-4	1	17	10	9	16	11	39
	50	31	43	38	84	33	65
	100	25	45	43	85	31	55
LeNet-5	1	2	7	16	16	13	7
	50	22	46	91	45	51	24
	100	23	49	100	46	53	28
RN-20	1	0	14	8	6	6	15
	50	0	58	62	40	44	57
	100	0	64	68	42	46	71
VGG-16	1	1	5	7	7	11	9
	50	3	48	36	34	39	44
	100	5	44	38	41	38	52
MobileNet	1	89	46	64	33	84	53
	50	400	783	880	709	1,198	872
	100	435	819	751	569	1,113	830
RN-50	1	7	11	11	6	22	15
	50	11,805	41,217	37,822	30,009	58,796	47,703
	100	11,793	41,132	37,810	29,979	58,747	47,712

and able to generate tests that capture a large scope of major functional behaviors of DNNs. For the case of VGG16 in Table 5, KMNC can more obviously distinguish accuracy of instances A and B with its generate tests compared with ones generated by the other criteria whose accuracy difference is about 1%. On other models, BNC, SNAC, TKNC, and BKNC perform well in many cases. Our in-depth investigation reveals the possible reason that the tests generated from these four criteria are more likely to be the error triggering tests for instance A or B.

The selection of the subject datasets and DNN models could be a threat to validity. In this paper, we try to counter this issue with 3 well-studied datasets with diverse complexity. For DNN models on MNIST and CIFAR-10, we follow the common machine learning training practice to obtain DNN models achieved with competitive performance. On ImageNet dataset, we select the well-pretrained models from Keras (ver.2.1.3) release. Another threat could be the randomness of the weight sampling in the mixed precision quantization, we counter this issue by repeating the same setting five times and averaging the results.

5 Related Work

In this section, we review the related work in the following three aspects: fuzz testing in traditional software, testing and verification of DNNs, and adversarial deep learning.

5.1 Fuzz Testing in Traditional Software

Fuzz testing has been widely used to safeguard software quality. Coverage guided grey-box fuzzing frameworks, such as AFL [15], libFuzzer [16], honggfuzz [17], and FOT [18] have been quite successful in detecting thousands of bugs in traditional software. On top of those, power scheduling [20] and some other techniques, such as [57], [58], have been demonstrated to be effective to guide the fuzzing procedure for different fuzzing purposes, such as increasing code coverage from low frequency execution traces or improving directedness for directed fuzzing scenarios.

On the other hand, several other methods have been proposed to improve the mutation quality by providing structure aware mutation strategies, including LangFuzz [59] and Skyfire [21] which generate or mutate the seeds according to some predefined grammars, or the libprotobuf mutator [60] which could be used to mutate protobuf supported formats. Compared to dumb mutators, more meaningful mutants can be generated to pass validity checks and detect more deeper bugs.

Other fuzzing techniques are also proposed to detect functionality bugs. For example, NEZHA [61] has been used to exploit the behavioral asymmetries between test programs to focus on inputs that are more likely to trigger logic bugs. The consistency of behaviors between different implementations serves as the oracle to detect the functionality bugs.

Finally, some works utilize machine learning or deep learning techniques to improve the effectiveness of fuzzing [62–67]. Different from these works, we attempt to perform fuzz testing on DNNs instead of leverage deep learning to fuzz traditional software.

5.2 Testing and Verification of DL Systems

Testing. DeepXplore [22] proposed a white-box differential testing technique to generate test inputs that potentially trigger inconsistencies between various DNNs; such inconsistencies may identify incorrect behaviors. They also investigated the usefulness of neuron coverage to measure how well the internal logic of a DNN is tested. In Tian and Pei *et al.* following work DeepTest [23], they further leveraged neuron coverage to guide testing of DNN-driven autonomous cars. DeepTest adopts the domain-specific metamorphic relations between the car behaviors across different input images to detect erroneous behaviors in a single DNN model, whereas DeepXplore [22] requires to check multiple DNNs.

DeepCover [68] adapted MC/DC test criteria [69] for DNNs, they showed its usefulness on small-scale neural networks (with no more than 500 neurons and 5 layers). Whether such criteria can scale to real-world sized DNN software remains to be investigated.¹¹ DeepGauge [44] generalized the concept of neuron coverage and proposed a set of 5 coverage criteria based on neuron numerical outputs. They have demonstrated that DeepGauge scales well to practical sized DNN models (*e.g.*, VGG-19, ResNet-50) and could capture erroneous behavior introduced by four state-of-the-art adversarial test

¹¹We have intended to include MC/DC criteria into *DeepHunter*. However, such coverage analysis on the large-scale seed batch in DNNs is computationally expensive, and we leave the efficient MC/DC integration in future work.

generation techniques (*i.e.*, FGSM, BIM, JSAM, and CW). DeepMutation [70], introduces a set of fault inject operators to generate mutant DNN models for test data quality evaluation. However, similar to traditional mutation testing, DeepMutation could be computationally intensive, since large amounts of mutant DNN models need to be generated, each of which is evaluated against the target test set. DeepCT performs combinatorial testing of DNN models to balance the huge input and latent space, and testing effectiveness [71]

Different from the existing works, this paper tends to propose a scalable and general-purpose coverage-guided fuzz testing framework for DNN software. We have integrated existing scalable coverage criteria into *DeepHunter* in order to guide testing and defect detection in large scale.

A concurrent work, TensorFuzz [72], tries to debug neural networks with coverage-guided fuzzing. *DeepHunter* differs from TensorFuzz mainly in three aspects. In TensorFuzz, the mutator provides only one type of mutation, which is additive noise, to the inputs. In *DeepHunter*, the mutator is entrusted with eight semantic-preserving metamorphic mutation types based on global and local image transformation, resulting in metamorphic inputs that are both diversified and plausible. Furthermore, in TensorFuzz, the feedback relies solely on one criterion, which is the basic neuron coverage. In *DeepHunter*, instead, we are employing a set of six multi-granularity neuron coverage criteria for providing multi-faceted feedback to the fuzzer. Most importantly, *DeepHunter* also differs from TensorFuzz with regards to the scope of measurement. In fact, the focus of our paper is to take a large-scale empirical study on multiple coverage to investigate their usefulness to guide test generation towards detect potential issues introduced during DNN development and deployment.

Verification. The reliability of DNNs has been investigated by recent work with formal guarantees [73–78]. Pulina *et al.* [74] proposed an abstraction-refinement approach to verify safety of a neural network with 6 neurons; Reluplex [75] adopted an SMT-based approach on a neural network with 300 ReLU nodes. DeepSafe [76] tried to identify safe regions in the input space using Reluplex as its core. A more recent work AI² [77] proposed an abstract interpretation technique to verify DNN software, through a well designed abstract domains and transformation operators. Since DNN software often handles high-dimensional input has large runtime internal states, designing more scalable and general verification methods towards complex real-world sized DNNs is challenging but important.

This paper further pushes quality assurance of DNNs from the automated fuzz testing perspective, by examining whether coverage guided fuzz testing could be a useful potential software quality assurance technique for DNN software. *DeepHunter* could scale to ImageNet like practical dataset with large DNN models like ResNet-50. Whether the potential integration of existing formal verification and fuzz testing is possible could be an interesting direction to explore.

5.3 Adversarial Deep Learning

A plethora of research has shown that carefully-crafted adversarial examples can undermine the robustness of DNNs [79–86]. In response to these attacks, several defenses have been proposed, such as ensemble method [87], GAN-based defense [88,89],

a certified approach [90], game theoretic based defense [91], stochastic quantization method [92], and so on [93–95]. However, it has been shown that none of these defenses or detection methods is robust enough against adaptive attacks [86].

Different from adversarial techniques, *DeepHunter* generates tests to detect both potential defects introduced during DNN development and quantization phase for deployment. In addition, these error triggering tests generated by *DeepHunter* are not limited to adversarial tests, which makes *DeepHunter* more general and promising than existing defenses or detection methods.

6 Conclusion and Future Work

Deep learning has seen tremendous success over the past decade, and has become the driving force for many novel intelligence applications. However, the quality assurance technique for DL is still at its early stage, and scalable DL testing framework is highly demanding. In this paper, we have proposed a coverage-guided fuzz testing framework for DNN software that systematically generates tests for detecting potential defects introduced during the DNN development and deployment phase. We have conducted a large-scale study to demonstrate its usefulness in facilitating defect detection, model quality evaluation, *etc.*, with data complexity increasing from MNIST to practical sized ImageNet.

Due to the computation resource limitation, we mainly focus on the investigation of how each single coverage criteria contributes to the effectiveness of *DeepHunter*, and on the opposite side, whether *DeepHunter* can effectively improve the corresponding coverage. How to intelligently combine the multiple criteria to further enhance the testing performance would be our future work. Furthermore, we intend to investigate how to integrate *DeepHunter* into DNN development and deployment practice, by providing useful feedback to help DL developers to enhance the DNN software quality. Since the investigation on quality assurance of deep learning is still at an early stage, we hope that *DeepHunter* can benefit both SE and AI communities, and facilitate further extensive studies towards constructing high quality DNN software.

References

1. J. E. Kelly III and S. Hamm, *Smart machines: IBM's Watson and the era of cognitive computing*. Columbia University Press, 2013.
2. Amazon, "Amazon Alexa," 2018. [Online]. Available: <https://developer.amazon.com/zh/alexa>
3. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
4. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, p. 484, 2016.
5. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

6. D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *NIPS*, 2012, pp. 2843–2851.
7. D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *CVPR*, 2012, pp. 3642–3649.
8. F. Zhang, J. Leitner, M. Milford, B. Upcroft, and P. Corke, "Towards vision-based deep reinforcement learning for robotic motion control," *arXiv:1511.03791*, 2015.
9. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.
10. F. Chollet *et al.*, "Keras," <https://github.com/fchollet/keras>, 2015.
11. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *NIPS 2017 Workshop Autodiff*, 2017.
12. A. Karpathy, "Software 2.0," <https://medium.com/@karpathy/software-2-0-a64152b37c35>, 2018.
13. Google Accident, "A Google self-driving car caused a crash for the first time," 2016. [Online]. Available: <https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report>
14. Uber Accident, "After Fatal Uber Crash, a Self-Driving Start-Up Moves Forward," 2018. [Online]. Available: <https://www.nytimes.com/2018/05/07/technology/uber-crash-autonomous-driveai.html>
15. "American Fuzzy Lop," 2018. [Online]. Available: <http://lcamtuf.coredump.cx/afl/>
16. "libFuzzer," 2018. [Online]. Available: <https://lvm.org/docs/LibFuzzer.html>
17. Google. (2018) honggfuzz. [Online]. Available: <https://github.com/google/honggfuzz>
18. H. Chen, Y. Li, B. Chen, Y. Xue, and Y. Liu, "Fot: A versatile, configurable, extensible fuzzing framework," in *The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 11 2018 (to appear).
19. S. Rawat, V. Jain, A. Kumar, L. Cojocar, C. Giuffrida, and H. Bos, "Vuzzer: Application-aware evolutionary fuzzing," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2017.
20. M. Böhme, V.-T. Pham, and A. Roychoudhury, "Coverage-based greybox fuzzing as markov chain," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1032–1043. [Online]. Available: <http://doi.acm.org/10.1145/2976749.2978428>
21. J. Wang, B. Chen, L. Wei, and Y. Liu, "Skyfire: Data-driven seed generation for fuzzing," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, May 2017, pp. 579–594. [Online]. Available: <https://doi.org/10.1109/SP.2017.23>
22. K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *Proceedings of the 26th Symposium on Operating Systems Principles*, 2017, pp. 1–18.
23. Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th International Conference on Software Engineering*. ACM, 2018, pp. 303–314.
24. A. Authors, "DeepGauge," <https://deepgauge.github.io/>, 2018.
25. R. Pressman, *Software Engineering: A Practitioner's Approach*, 7th ed. New York, NY, USA: McGraw-Hill, Inc., 2010.
26. N. B. Ruparelia, "Software development lifecycle models," *SIGSOFT Softw. Eng. Notes*, vol. 35, no. 3, pp. 8–13, May 2010.

27. L. Ma, F. Juefei-Xu, M. Xue, Q. Hu, S. Chen, B. Li, Y. Liu, J. Zhao, J. Yin, and S. See, "Secure Deep Learning Engineering: A Software Quality Assurance Perspective," *ArXiv e-prints*, Oct. 2018.
28. B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *CVPR*, June 2018.
29. S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *ICLR*, 2015, pp. 1737–1746.
30. M. Courbariaux, Y. Bengio, and J.-P. David, "Training deep neural networks with low precision multiplications," *arXiv preprint arXiv:1412.7024*, 2014.
31. S. Wu, G. Li, F. Chen, and L. Shi, "Training and inference with integers in deep neural networks," *arXiv preprint arXiv:1802.04680*, 2018.
32. M. Courbariaux and Y. Bengio, "Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1," *arXiv preprint arXiv:1602.02830*, 2016.
33. M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *NIPS*, 2015, pp. 3105–3113.
34. M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *ECCV*, 2016, pp. 525–542.
35. C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," *ICLR*, 2017.
36. I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.
37. F. Juefei-Xu, V. N. Boddeti, and M. Savvides, "Perturbative neural networks," in *CVPR*. IEEE, June 2018, pp. 3310–3318.
38. F. Juefei-Xu, V. Boddeti, and M. Savvides, "Local binary convolutional neural networks," in *CVPR*. IEEE, July 2017, pp. 19–28.
39. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *ICLR*, 2016.
40. "NVIDIA TensorRT," 2018. [Online]. Available: <https://developer.nvidia.com/tensorrt>
41. "Peach Fuzzer Platform," 2018. [Online]. Available: <https://www.peach.tech/products/peach-fuzzer/peach-platform/>
42. C. Cadar, D. Dunbar, D. R. Engler *et al.*, "Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs." in *OSDI*, vol. 8, 2008, pp. 209–224.
43. S. Gan, C. Zhang, X. Qin, X. Tu, K. Li, Z. Pei, and Z. Chen, "Collafl: Path sensitive fuzzing," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 679–696.
44. L. Ma, F. Juefei-Xu, J. Sun, C. Chen, T. Su, F. Zhang, M. Xue, B. Li, L. Li, Y. Liu *et al.*, "Deepgauge: Multi-granularity testing criteria for deep learning systems," *The 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE 2018)*, 2018.
45. M. Zalewski. (2014) Technical "whitepaper" for afl-fuzz. [Online]. Available: http://lcamtuf.coredump.cx/afl/technical_details.txt
46. F. Chollet *et al.*, "Keras applications," <https://keras.io/applications/>, 2018.
47. Y. LeCun and C. Cortes, "The MNIST database of handwritten digits," 1998.
48. N. Krizhevsky, H. Vinod, C. Geoffrey, M. Papadakis, and A. Ventresque, "The cifar-10 dataset," <http://www.cs.toronto.edu/kriz/cifar.html>, 2014.
49. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
50. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
51. C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating Adversarial Examples with Adversarial Networks," *ArXiv e-prints*, Jan. 2018.

52. N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Security and Privacy (SP), IEEE Symposium on*, 2017, pp. 39–57.
53. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
54. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
55. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
56. I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016, vol. 1.
57. M. Böhme, V.-T. Pham, M.-D. Nguyen, and A. Roychoudhury, "Directed greybox fuzzing," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: ACM, 2017, pp. 2329–2344. [Online]. Available: <http://doi.acm.org/10.1145/3133956.3134020>
58. Y. L. B. C. X. X. W. Hongxu Chen, Yinxing Xue and Y. Liu, "Hawkeye: Towards a desired directed grey-box fuzzer," in *Proceedings of the 25th ACM Conference on Computer and Communications Security*. ACM, 2018.
59. C. Holler, K. Herzig, and A. Zeller, "Fuzzing with code fragments," in *Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)*. Bellevue, WA: USENIX, 2012, pp. 445–458. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/holler>
60. Google. (2017) libprotobuf-mutator. [Online]. Available: <https://github.com/google/libprotobuf-mutator>
61. T. Petsios, A. Tang, S. Stolfo, A. D. Keromytis, and S. Jana, "Nezha: Efficient domain-independent differential testing," in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 615–632.
62. P. Godefroid, H. Peleg, and R. Singh, "Learn&fuzz: Machine learning for input fuzzing," *CoRR*, vol. abs/1701.07232, 2017. [Online]. Available: <http://arxiv.org/abs/1701.07232>
63. Payatu. (2018) Cloudfuzz: Machine learning powered content specific input generation for fuzzing.
64. G. Yan, J. Lu, Z. Shu, and Y. Kucuk, "Exploitmeter: Combining fuzzing with machine learning for automated evaluation of software exploitability," in *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, Aug 2017, pp. 164–175.
65. M. Rajpal, W. Blum, and R. Singh, "Not all bytes are equal: Neural byte sieve for fuzzing," *CoRR*, vol. abs/1711.04596, 2017. [Online]. Available: <http://arxiv.org/abs/1711.04596>
66. C. Cummins, P. Petoumenos, A. Murray, and H. Leather, "Compiler fuzzing through deep learning," in *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA 2018. New York, NY, USA: ACM, 2018, pp. 95–105. [Online]. Available: <http://doi.acm.org/10.1145/3213846.3213848>
67. D. She, K. Pei, D. Epstein, J. Yang, B. Ray, and S. Jana, "NEUZZ: efficient fuzzing with neural program learning," *CoRR*, vol. abs/1807.05620, 2018. [Online]. Available: <http://arxiv.org/abs/1807.05620>
68. Y. Sun, X. Huang, and D. Kroening, "Testing Deep Neural Networks," *ArXiv e-prints*, Mar. 2018.
69. H. Kelly J., V. Dan S., C. John J., and R. Leanna K., "A practical tutorial on modified condition/decision coverage," NASA, Tech. Rep., 2001.
70. L. Ma, F. Zhang, J. Sun, M. Xue, B. Li, F. Juefei-Xu, C. Xie, L. Li, Y. Liu, J. Zhao *et al.*, "Deepmutation: Mutation testing of deep learning systems," *The 29th IEEE International Symposium on Software Reliability Engineering (ISSRE)*, 2018.
71. L. Ma, F. Zhang, M. Xue, B. Li, Y. Liu, J. Zhao, and Y. Wang, "Combinatorial testing for deep learning systems," *arXiv preprint arXiv:1806.07723*, 2018.

72. A. Odena and I. Goodfellow, “Tensorfuzz: Debugging neural networks with coverage-guided fuzzing,” *arXiv preprint arXiv:1807.10875*, 2018.
73. M. Wicker, X. Huang, and M. Kwiatkowska, “Feature-guided black-box safety testing of deep neural networks,” *CoRR*, vol. abs/1710.07859, 2017. [Online]. Available: <http://arxiv.org/abs/1710.07859>
74. L. Pulina and A. Tacchella, “An abstraction-refinement approach to verification of artificial neural networks,” in *International Conference on Computer Aided Verification*. Springer, 2010, pp. 243–257.
75. G. Katz, C. W. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, “Reluplex: An efficient SMT solver for verifying deep neural networks,” *CoRR*, vol. abs/1702.01135, 2017. [Online]. Available: <http://arxiv.org/abs/1702.01135>
76. D. Gopinath, G. Katz, C. S. Pasareanu, and C. Barrett, “Deepsafe: A data-driven approach for checking adversarial robustness in neural networks,” *CoRR*, vol. abs/1710.00486, 2017. [Online]. Available: <http://arxiv.org/abs/1710.00486>
77. D. D.-C. P. T. S. C. M. V. Timon Gehr, Matthew Mirman, “Ai2: Safety and robustness certification of neural networks with abstract interpretation,” in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018.
78. K. Pei, Y. Cao, J. Yang, and S. Jana, “Towards practical verification of machine learning: The case of computer vision systems,” *CoRR*, vol. abs/1712.01785, 2017. [Online]. Available: <http://arxiv.org/abs/1712.01785>
79. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *ICLR*, 2015.
80. W. He, B. Li, and D. Song, “Decision boundary analysis of adversarial examples,” in *ICLR*, 2018.
81. W. Brendel, J. Rauber, and M. Bethge, “Decision-based adversarial attacks: Reliable attacks against black-box machine learning models,” in *ICLR*, 2018.
82. Z. Zhao, D. Dua, and S. Singh, “Generating natural adversarial examples,” in *ICLR*, 2018.
83. C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, “Spatially transformed adversarial examples,” in *ICLR*, 2018.
84. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *ICLR*, 2014.
85. N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *IEEE Symposium on Security and Privacy, 2017*, 2017.
86. ———, “Adversarial examples are not easily detected: Bypassing ten detection methods,” in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 3–14.
87. F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” in *ICLR*, 2018.
88. P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-gan: Protecting classifiers against adversarial attacks using generative models,” in *ICLR*, 2018.
89. Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, “Pixeldefend: Leveraging generative models to understand and defend against adversarial examples,” in *ICLR*, 2018.
90. A. Raghunathan, J. Steinhardt, and P. Liang, “Certified defenses against adversarial examples,” in *ICLR*, 2018.
91. G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar, “Stochastic activation pruning for robust adversarial defense,” in *ICLR*, 2018.
92. A. Galloway, G. W. Taylor, and M. Moussa, “Attacking binarized neural networks,” in *ICLR*, 2018.
93. C. Guo, M. Rana, M. Cisse, and L. van der Maaten, “Countering adversarial images using input transformations,” in *ICLR*, 2018.

94. J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, “Thermometer encoding: One hot way to resist adversarial examples,” in *ICLR*, 2018.
95. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *ICLR*, 2018.