MTFuzz: Fuzzing with a Multi-Task Neural Network

Dongdong She  
Columbia University  
New York, USA  
ds3619@columbia.edu

Rahul Krishna  
Columbia University  
New York, USA  
rk3080@columbia.edu

Lu Yan  
Shanghai Jiao Tong University  
Shanghai, China  
jiaodayanlu@sjtu.edu.cn

Suman Jana  
Columbia University  
New York, USA  
suman@cs.columbia.edu

Baishakhi Ray  
Columbia University  
New York, USA  
rayb@cs.columbia.edu

ABSTRACT

Fuzzing is a widely used technique for detecting software bugs and vulnerabilities. Most popular fuzzers generate new inputs using an evolutionary search to maximize code coverage. Essentially, these fuzzers start with a set of seed inputs, mutate them to generate new inputs, and identify the promising inputs using an evolutionary fitness function for further mutation. Despite their success, evolutionary fuzzers tend to get stuck in long sequences of unproductive mutations. In recent years, machine learning (ML)-based mutation strategies have reported promising results. However, the existing ML-based fuzzers are limited by the lack of quality and diversity of the training data. As the input space of the target programs is high dimensional and sparse, it is prohibitively expensive to collect many diverse samples demonstrating successful and unsuccessful mutations to train the model.

In this paper, we address these issues by using a Multi-Task Neural Network that can learn a compact embedding of the input space based on diverse training samples for multiple related tasks (i.e., predicting different types of coverage). The compact embedding can be used to guide the mutation process effectively by focusing most of the mutations on the parts of the embedding where the gradient is high. Our results show that MTFuzz uncovers 11 previously unseen bugs and achieves an average of 2x more edge coverage compared with 5 state-of-the-art fuzzers on 10 real-world programs.

CCS CONCEPTS

• Software and its engineering → Software testing and debugging.

KEYWORDS

Fuzzing, Neural Networks, gradient-guided optimization

1 INTRODUCTION

Coverage-guided graybox fuzzing is a widely used technique for detecting bugs and security vulnerabilities in real-world software [7, 26, 27, 30, 42, 46, 52, 53, 64, 71, 74, 76, 76]. The key idea behind a fuzzer is to execute the target program on a large number of automatically generated test inputs and monitor the corresponding executions for buggy behaviors. However, as the input spaces of real-world programs are typically very large, unguided test input generation is not effective at finding bugs. Therefore, most popular graybox fuzzers use evolutionary search to generate new inputs; they mutate a set of seed inputs and retain only the most promising inputs (i.e., inputs exercising new program behavior) for further mutation [33, 38, 42, 43, 53, 71, 73, 74, 76].

However, the effectiveness of traditional evolutionary fuzzers tends to decrease significantly over fuzzing time. They often get stuck in long sequences of unfruitful mutations, failing to generate inputs that explore new regions of the target program [20, 63, 64]. Several researchers have worked on designing different mutation strategies based on various program behaviors (e.g., focusing on rare branches, call context, etc.) [20, 42]. However, program behavior changes drastically, not only across different programs but also across different parts of the same program. Thus, finding a generic robust mutation strategy still remains an important open problem.

Recently, Machine Learning (ML) techniques have shown initial promise to guide the mutations [59, 61, 64]. These fuzzers typically use existing test inputs to train ML models and learn to identify promising mutation regions that improve coverage [33, 59, 61, 64]. Like any other supervised learning technique, the success of these models relies heavily on the number and diversity of training samples. However, collecting such training data for fuzzing that can demonstrate successful/unsuccessful mutations is prohibitively expensive due to two main reasons. First, successful mutations that increase coverage are often limited to very few, sparsely distributed input bytes, commonly known as hot-bytes, in a high-dimensional input space. Without knowing the distribution of hot-bytes, it is extremely hard to generate successful mutations over the sparse, high-dimensional input space [61, 64]. Second, the training data must be diverse enough to expose the model to various program behaviors that lead to successful/unsuccessful mutations—this is also challenging as one would require a large number of test cases exploring different program semantics. Thus, the ML-based fuzzers suffer from both sparsity and lack of diversity of the target domain.
In this paper, we address these problems using Multi-Task Learning, a popular learning paradigm used in domains like computer vision to effectively learn common features shared across related tasks from limited training data. In this framework, different participating tasks allow an ML model to effectively learn a compact and more generalized feature representation while ignoring task-specific noises. To jointly learn a compact embedding of the inputs, in our setting, we use different tasks for predicting the relationship between program inputs and different aspects of fuzzing-related program behavior (e.g., different types of edge coverage). Such an architecture addresses both the data sparsity and lack of diversity problem. The model can simultaneously learn from diverse program behaviors from different tasks as well as focus on learning the important features (hot bytes in our case) across all tasks. Each participating task will provide separate pieces of evidence for the relevance or irrelevance of the input features [60].

To this end, we design, implement, and evaluate MTFuzz, a Multi-task Neural Network (MTNN) based fuzzing framework. Given the same set of test inputs, MTFuzz learns to predict three different code coverage measures showing various aspects of dynamic program behavior: (i) edge coverage: which edges are explored by a test input [64, 76]? (ii) approach-sensitive edge coverage: if an edge is not explored, how far off it is (i.e., approach level) from getting triggered [5, 47, 48, 55]? and (iii) context-sensitive edge coverage: from which call context an explored edge is called [20, 72]? Note that our primary task, like most popular fuzzers, is still generating mutants that can increase edge coverage. However, the use of call context and approach level provides additional information to boost input-output diversity.

In particular, the underlying MTNN contains a group of hidden layers shared across the participating tasks, while still maintaining task-specific output layers. The last shared layer learns a compact embedding of the input space as shown in Figure 1. Such an embedding captures a generic compressed representation of the inputs while preserving the important features, i.e., hot-byte distribution. We further compute a saliency score [63] of each input byte by computing the gradients of the embedded representation w.r.t. the input byte. Saliency scores are often used in computer vision models to identify the important features by analyzing the importance of that feature w.r.t. an embedded layer [66]. By contrast, in this paper, we use such saliency scores to guide the mutation process—focus the mutations on bytes with high saliency scores.

Our MTNN architecture also allows the compact embedding layer, once trained, to be transferred across different programs that operate on similar input formats. For example, compact-embedding learned with MTFuzz for one xml parser may be transferred to other xml parsers. Our results (RQ4) show that such transfer is quite effective and it mitigates the need to generate high quality data from scratch for new programs which can be quite expensive.

We evaluate MTFuzz on 10 real world programs against 5 state-of-the-art fuzzers. MTFuzz covers at least 1000 more edges on 5 programs and several 100 more on the rest. MTFuzz also finds a total of 71 real-world bugs (11 previously unseen) (see RQ1). When compared to learning each task individually, MTFuzz offers significantly more edge coverage (see RQ2). Lastly, our results from transfer learning show that the compact-embedding of MTFuzz can be transferred within parsers for xml and elf binaries.

Overall, our paper makes the following key contributions:

- We present a novel fuzzing framework based on multi-task neural networks called MTFuzz that learns a compact embedding of otherwise sparse and high-dimensional program input spaces. Once trained, we use the salience score of the embedding layer outputs w.r.t. the input bytes to guide the mutation process.
- Our empirical results demonstrate that MTFuzz is significantly more effective than current state-of-the-art fuzzers. On 10 real world programs, MTFuzz achieves an average of 2× and up to 3× edge coverage compared to Neuzz, the state-of-the-art ML-based fuzzer. MTFuzz also finds 11 previously unknown bugs other fuzzers fail to find. We are currently working on open-sourcing our tool and reporting the bugs to the developers.
- Our results also demonstrate that transferring the compact embedding across programs expecting similar input formats can significantly increase fuzzing efficiency. For example, transferred embeddings for different file formats like ELF and XML can achieve up to 12× more edge coverage compared to existing fuzzers.
- We offer an open-source replication package to reproduce and extend MTFuzz at https://github.com/ARISe-Lab/MTFuzz.

2 BACKGROUND: MULTI-TASK NETWORKS

Multi-task Neural Networks (MTNN) are becoming increasingly popular in many different domains including optimization [6, 34], natural language processing [12, 23], and computer vision [67]. The key intuition behind MTNN is that it is useful for related tasks to be learned jointly so that each task can benefit from the relevant information available in other tasks [17, 18, 67, 77]. For example, if
we learn to ride a unicycle, a bicycle, and a tricycle simultaneously, experiences gathered from one usually help us to learn the other tasks better [78]. More formally, the objective of an MTNN may be defined as follows:

In this paper, we use a popular MTNN architecture called hard parameter sharing [18], which contain two groups of layers (see Fig. 5): a set of initial layers shared among all the tasks, and several individual task-specific output layers. The shared layers enable a MTNN to find a common feature representation across all the task. The task specific layers use the shared feature representation to generate predictions for individual tasks [39, 60, 67].

**MTNN Training.** While MTNNs can be used in many different ML paradigms, in this paper we primarily focus on supervised learning. We assume that the training process has access to a training dataset \(X = \{x_1, x_2, ..., x_n\}\). The training data also contains the ground truth output labels for each task and \(n\) training sample pair. We train the MTNN on the training data using standard back-propagation to minimize a multi-task loss as described below.

**Multi-task Loss.** Training an MTNN is guided by a multi-task loss function, \(L\). We assume that each individual task \(\tau_i\) in the set of tasks \(\tau = \{\tau_1, \tau_2, ..., \tau_m\}\) has a corresponding loss function \(L_i\). The multi-task loss is computed as a weighted sum of each individual task loss. More formally, it is given by \(L = \sum_{i=1}^{m} \alpha_i \cdot L_i\). Here, \(\alpha_i\) represents the weight assigned to task \(i\). The goal of training is to reduce the overall loss. In practice, the actual values of the weights are decided based on the relative importance of each task. Most existing works assign equal weights to the tasks [44, 69, 70].

The multi-task loss function forces the shared layer to learn a general input representation for all tasks offering two benefits:

- **Increased generalisability.** Prior results have demonstrated that the overall risk of overfitting in multi-task models is reduced by an order of \(m\) (\(m\) is the number of tasks) compared to single task models [10]. Intuitively, the more tasks an MTNN has to learn simultaneously, the more general the required representation should be to capture the important features of all the tasks. This would prevent the representation from overfitting to task-specific features.

- **Reduced sparsity.** The shared embedding layer in an MTNN can be designed to increase the compactness of the learned input representation. A shared embedding layer with far fewer nodes compared to the input layer can be trained to retain the same expressive power of the input with respect to a given set of tasks. In such compact embedding, the important features across different tasks will be boosted with each task contributing its own set of relevant features [60].

## 3 METHODOLOGY

This section presents a brief overview of MTFuzz that aims to maximize edge coverage with the aid of two additional coverage measures: context-sensitive edge coverage and approach-sensitive edge coverage using multi task learning. Fig. 1 illustrates an end-to-end workflow of the proposed approach. The first stage trains an MTNN to produce a compact embedding of an otherwise sparse input space while preserving information about the input bytes that have the highest likelihood to impact code coverage (§3.2). The second stage identifies these hot bytes and focuses on mutating them (§3.3). Finally, in the third stage, the seed corpus is updated to retain only the most interesting new inputs (§3.4).

### 3.1 Modeling Coverage as Multiple Tasks

The goal of any ML-based fuzzers, including MTFuzz, is to learn a mapping between input space and code coverage. The most common coverage explored in the literature is edge coverage, which is a reasonably effective measure and quite easy to instrument. However, it tends to be too coarse-grained missing many interesting program behavior (e.g. explored call context) that are known to be important to fuzzing. One workaround can be to model path coverage by tracking the program execution path per input. However, keeping track of all the explored paths can be computationally intractable since it can quickly lead to a state-space explosion for large programs [72]. As an alternative, in this work, we propose a middle ground: we model the edge coverage as the primary task of the MTNN, while choosing two other finer granular coverage metrics (approach-sensitive edge coverage and context-sensitive edge coverage) as auxiliary tasks to provide useful additional context to edge coverage while still preventing state-space explosion.

#### 3.1.1 Edge coverage: Primary Task

Edge coverage measures how many unique control-flow edges are triggered by a test input as it interacts with the program. It has become the de-facto code coverage metric [42, 64, 74, 76] for fuzzing. We model edge coverage prediction as the primary task of our multi-task network, which takes a binary test case as input and predicts (i.e. outputs) the edges that could be covered by the test case. For each input, we represent the edge coverage as *edge bitmap*, where value per edge is set to 1 or 0 depending on whether the edge is exercised by the input.

In particular, in the control-flow-graph of a program, an edge connects two basic blocks (denoted by prev_block and cur_block) [76]. A unique edge_id is obtained as: hash(prev_block, cur_block). For each edge_id, there is a bit allocated in the bitmap. For every input, the edge_ids in the corresponding edge bitmap are set to 1 or 0, depending on whether or not those edges were triggered.

#### 3.1.2 Approach-Sensitive Edge Coverage: Auxiliary Task 1

For an edge that is not exercised by an input, we measure how far off the edge is from getting triggered. Such a measure provides additional contextual information of an edge. For example, if two test inputs failed to trigger an edge, however one input reached “closer” to the unexplored edge than the other, traditional edge coverage would treat both inputs the same. However, using a proximity measure, we can discern between the two inputs and mutate the closer input so that it can reach the unexplored edge.

![Figure 2: Approach Bitmap vs. Edge Bitmap](image-url)

*Figure 2: Approach Bitmap vs. Edge Bitmap. The edge ‘d’ has a visited parent edge ‘b’ and is thus marked 0.5 in the approach bitmap.*
To this end, approach-sensitive edge coverage extends edge coverage by offering a distance measure that computes the distance between an unreached edge and the nearest edge triggered by an input. This is a popular measure in the search-based software engineering literature [5, 47, 48], where instead of assigning a binary value (0 or 1), as in the case of edge coverage, approach level assigns a numeric value between 0 and 1 to represent the edges [49]; If an edge is triggered, it is assigned 1. However, if the edge is not triggered, but one of its parents is triggered, then the non-triggered edge is assigned a value of β (we use β = 0.5). If neither the edge nor its parents are triggered, it is assigned 0. This is illustrated in Fig. 2. Note that, for a given edge, we refrain from using additional ancestors farther up the control-flow graph to limit the computational burden. The approach sensitive coverage is represented in an approach bitmap, where per unique edge_id, we set its approach level value, as shown in Fig. 2.

We model this metric in our Multi-task Neural Network framework as an auxiliary task, where the task takes binary test cases as inputs and learn to predict the corresponding approach-level bitmap.

3.1.3 Context-sensitive Edge Coverage: Auxiliary Task 2. Edge coverage cannot distinguish between two different test inputs trigger the same edge, but via completely different internal states (e.g., through the same function called from different sites in the program). This distinction is valuable because, reaching an edge via a new internal state (e.g., through a new function call site) may trigger a vulnerability hidden deep within the program logic. Augmenting edge coverage with context information regarding internal states of the program may help alleviate this problem [20].

Consider the example in Fig. 3. Here, for an input [1, 0], the first call to function Foo() appears at site line 12 and it triggers the if condition (on line 2); the second call to Foo() appears on site line 11 and it triggers the else condition (on line 5). As far as edge coverage is concerned, both the edges of the function Foo() (on lines 2 and 5) have been explored and any additional inputs will remain uninteresting. However, if we provide a new input say [0, 8], we would first trigger line 5 of Foo when it is called from line 12. Then we trigger line 2 of Foo from line 14 and further cause a buffer overflow at line 3 because a 12 bytes string is written into a 8 bytes destination buffer buf. Moreover, the input [0, 8] will not be saved by edge coverage fuzzier since it triggers no new edges. Frequently called functions (like strncpy()) may be quite susceptible such crashes [72].

In order to overcome this challenge, Chen et al. [20] propose keeping track of the call stack in addition to the edge coverage by maintaining tuple: (call stack, prev block, cur block). Fig. 4 shows the additional information provided by context-sensitive edge coverage over edge coverage. Here, we see that an example where a buggy input [0, 8] has the exact same edge coverage as the clean input [1, 0]. However, the call context information can differentiate these two inputs based on the call stacks at lines 12 and 14.

We model context-sensitive edge coverage in our framework as an auxiliary task. We first assign a unique id to every call. Next, at run time, when we encounter a call at an edge (edge_id), we first compute the hash value with @ to record all the functions on current call stack as: call_stack = call_id; @ ... @ call_id_n, where call_id_i represents the i-th function on current call stack.

```c
void foo(char* addr, int a) {
    if (a > 0){
        strncpy(addr, “I might overflow.”, a+4);
        return;
    } else
        return;
}

int main(int argc, char** input)
{
    char buf[8];
    ...
    foo(buf, input[0]);
    ...
    foo(buf, input[1]);
    ...
}
```

Figure 3: An example C-code to demonstrate the usefulness of using context-sensitive measures. Measures such as edge coverage will fail to detect a possible bug in strncpy().

Next, we compute the context sensitive edge id as: call_trace_id = call_stack @ edge_id.

Thus we obtain a unique call_trace_id for every function called from different contexts (i.e., call sites). We then create a bit-map of all the call_trace_ids. Unlike existing implementations of context-sensitive edge coverage [20, 72], we assign an additional id to each call instruction while maintaining the original edge_id intact. Thus, the total number of elements in our bit map reduces to sum of call_trace_ids and edge_ids rather than a product of call_trace_ids and edge_ids. An advantage of our design is that we minimize the bitmap size. In practice, existing methods requires around 7× larger bitmap size than just edge coverage [20]; our implementation only requires around 1.3× bitmap size of edge coverage. The smaller bitmap size can avoid edge explosion and further boost fuzzing performance.

In our multi-tasking framework, the context-sensitive edge coverage “task” is trained to predict the mapping between the inputs and the corresponding call_trace_ids bitmaps. This can enable us to learn the difference between two inputs in a more granular fashion. For example, an ML model can learn that under certain circumstances, the second input byte (input[1] in Fig. 3) can cause crashes. This information cannot be learned by training to predict for edge coverage alone since both inputs will have the same edge coverage (as shown in Fig. 4).

3.2 Stage-I: Multi-Task Training

This phase builds a multi-task neural network (MTNN) that can predict different types of edge coverage given a test input. The
Figure 5: The MTNN architecture representing the n-dimensional input layer $xb^i \in xb$; m-dimensional compact embedding layer $zb^i \in z$, s.t. $m < n$, with a function $F(\cdot)$ to map input $xb$ and the embedding layer $zb$; three task-specific layers.

A trained model is designed to produce a more general and compact embedding of the input space focusing only on those input bytes that are most relevant to all the tasks. This compact representation will be reused by the subsequent stages of the program to identify the most important bytes in the input (i.e., the hot-bytes) and guide mutations on those bytes.

### 3.2.1 Architecture

Fig. 5 shows the architecture of the MTNN. The model contains an encoder (shared among all the tasks) and three task-specific decoders. The model takes existing test input bytes as input and outputs task-specific bitmap. Each input byte corresponds to one input node, and each bitmap value corresponds to an output node.

**Encoder** Comprises of one input layer followed by three progressively narrower intermediate layers. The total number of nodes in the input layer is equal to the total number of bytes in the largest input in the seed corpus. All shorter inputs are padded with 0x00 for consistency. The last layer of the encoder is a compact representation of the input to be used by all the tasks (green in Fig. 5).

**Decoders.** There are three task-specific decoders (shown in lilac in Fig. 5). Each task-specific decoder consists of three intermediate layers that grow progressively wider. The last layer of each of the decoder is the output layer. For edge coverage task, there is one node in the output layer for each unique edge_id, likewise for context-sensitive edge coverage there is one output node for each call_trace_id, and for approach-sensitive edge coverage there is one output node for each unique edge_id but they take continuous values (see Fig. 2).

### 3.2.2 Loss functions

The loss function of a MTNN is a weighted sum of the task-specific loss functions. Among our three tasks, edge coverage and context-sensitive edge coverage are modeled as classification tasks and approach-sensitive edge coverage is modeled as a regression task. Their loss functions are designed accordingly.

**Loss function for approach-sensitive edge coverage.** Approach-level measures how close an input was from an edge that was not triggered. This distance is measured using a continuous value between 0 and 1. Therefore, this is a regression problem and we use mean squared error loss, given by $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, where $Y_i$ is the prediction and $\hat{Y}_i$ is the ground truth.

**Loss functions for edge coverage and context-sensitive edge coverage.** The outputs of both these tasks are binary values where 1 means an input triggered the edge_id or the call_trace_id and 0 otherwise. We find that while some edge_ids or call_trace_ids are triggered very rarely creating an imbalance. This usually happens when an input triggers a previously unseen (rare) edges. Due to this imbalance, training with an off-the-shelf loss functions such as cross entropy is ill suited because it causes a lot of false negative errors often missing these rare edges.

To address this issue, we introduce a parameter called penalty (denoted by $\beta$) to penalize these false negatives. The penalty is the ratio of the number of times an edge is not invoked over the number of times it is invoked. That is,

$$\text{Penalty} = \frac{\# \text{times an edge_id/call_trace_id is not invoked}}{\# \text{times an edge_id/call_trace_id is invoked}}$$

Here, $\beta_i$ represents the penalty for every applicable task $i \in T$ and it is dynamically evaluated as fuzzing progresses. Using $\beta_i$, we define an adaptive loss for classification tasks in our MTNN as:

$$\mathcal{L}_T = - \sum_{\text{edge}} (\beta_i \cdot p \cdot \log(\hat{p}) + (1 - p) \cdot \log(1 - \hat{p}))$$

In Eq. 1, $\mathcal{L}_T$ results in two separate loss functions for edge coverage and context-sensitive edge coverage. The penalty ($\beta_i$) is used to penalize false-positive and false-negative errors. $\beta_i > 1$ penalizes $p \cdot \log(\hat{p})$, representing false positives; $\beta_i = 1$ penalizes both false positives and false negatives equally; and $\beta_i < 1$ penalizes $(1 - p) \cdot \log(1 - \hat{p})$, representing false positives.

### 3.3 Stage-II: Guided Mutation

This phase uses the trained MTNN to generate new inputs that can maximize the code coverage. This is achieved by focusing mutation on the byte locations in the input that can most influence the branching behavior of the program (hot-bytes).

We use the compact embedding layer of the MTNN (shown in green in Fig. 5) to infer the hot-byte distribution. The compact embedding layer is well suited for this because it (a) captures the most semantically meaningful features (i.e., bytes) in the input in a compact manner; and (b) learns to ignore task-specific noise patterns [41] paying more attention to the important bytes that apply to all tasks [11, 21, 40].

Formally, we can represent the input (shown in orange in Fig. 5) as a byte vector $xb = (xb^1, xb^2, \ldots, xb^n) \in [0, 255]^n$, where $xb^i$ is the $i^{th}$ byte in $xb$ and $n$ represents the input dimensions (i.e., number of bytes). Then, after the MTNN has been trained, we obtain the compact embedding layer $z = (z^1, z^2, \ldots, z^m)$ consisting of $m$ nodes.

We note that for every byte that changes in $xb = (xb^1, \ldots, xb^n)$, we obtain corresponding changes to every node in $z = (z^1, \ldots, z^m)$. The amount of the change is determined by how influential each of the $n$ bytes in the input are to all the tasks in the MTNN model. Changes to the hot-bytes, which are more influential, will result in larger changes to $z$. We use this property to discover the hot-bytes.

To determine how influential each of the bytes in $xb$ are, we compute the partial derivatives of the nodes in compact layer with respect to all the input bytes. The partial derivative of the $j^{th}$ node
in the embedding layer with respect to the $i$-th input byte is:

$$\nabla_{xb} z = \frac{\partial f_i(x)}{\partial x_{bi}} = \left| \frac{\partial f_i}{\partial x_{bi}} \right|_{i=1,..,n; j=1,..,m}$$

(2)

In order to infer the importance of each byte $x_{bi} \in xb$, we define a saliency score for each byte, denoted by $S(x_{bi})$. We compute the saliency score as follows:

$$S(x_{bi}) = \sum_{j=1}^{m} \left| \frac{\partial f_i}{\partial x_{bi}} \right|_{i=1,..,n}$$

(3)

The saliency score $S(x_{bi})$ is the sum of all the partial derivatives in $\nabla_{xb} z$ w.r.t. to the byte $x_{bi}$. The numeric value of each of the $n$ elements in $S(xb)$ determines the hotness of each of the bytes. The larger the saliency score of an input byte $x_{bi}$ the more likely it is to be a hot-byte. Using the saliency score ($S(x_{bi})$) we can now mutate the existing inputs to generate new ones. To do this, we identify the top $k$ bytes with the largest saliency values—these are the byte-locations that will be mutated by our algorithm.

For each selected byte-locations, we create a new mutated input by changing the bytes to all permissible values between 0 and 255. Since this only happens to the top $k$ bytes, the number of newly mutated seeds remains manageable. We use these mutated inputs for fuzzing and monitor various coverage.

3.4 Stage-III: Seed Selection & Incremental Learning

In this step, MTFuzz samples some of the mutated inputs from the previous stage and further retrain the model. Sampling inputs is an important step as the choice of inputs can significantly affect the fuzzing performance. Also, as fuzzing progresses, the pool of available inputs keeps growing in size. Without some form of sampling strategy, training the NN and computing gradients would take prohibitively long.

To this end, we propose an importance sampling [54] strategy where inputs are sampled such that they reach some important region of the control-flow graph instead of randomly sampling from available input. In particular, our sampling strategy first retains all inputs that invoke previously unseen edges. Then, we sort all the seen edges by their rarity. The rarity of an edge is computed by counting how many inputs trigger that specified edge. Finally, we select the $T$-most rarest edges and include at least one input triggering each of these rare edges. We reason that, by selecting the inputs that invoke the rare edges, we may explore deeper regions of the program on subsequent mutations. In order to limit the number of inputs sampled, we introduce a sampling budget $K$ that determines how many inputs will be selected per iteration.

Using these sampled inputs, we retrained the model periodically to refine its behavior—as more data is becoming available about new execution behavior, retraining makes sure the model has knowledge about them and thus, can make more informed predictions.

3.5 Implementation

Our NN model is implemented in Keras-2.2.3 [22] with Tensorflow-1.8.0 [4] as a backend. The NN is based on a feed-forward model, composed of one shared encoder and three independent decoders. The encoder represents an input file into a 512 compact feature vector and feed it into three following decoders to perform different task predictions. For encoder, we use three hidden layers with dimensions 2048, 1024 and 512. For each decoder, we use one final output layer to perform corresponding task prediction. The dimension of final output layer is determined by different programs. We use ReLU as activation function for all hidden layers. We use sigmoid as the activation function for the output layer. The NN model is trained for 100 epochs to achieve high test accuracy (around 95% on average). We use Adam optimizer and learning rate is 0.001.

4 EXPERIMENTAL RESULTS

4.1 Study Subjects

We evaluate MTFuzz on 10 real-world programs, as shown in Table 2. To demonstrate the performance of MTFuzz, we compare the edge coverage and number of bugs detected by MTFuzz with 5 state-of-the-art fuzzers listed in Tab. 1. Training Data Collection. All our measurements are performed on a system running Ubuntu 18.04 with Intel Xeon E5-2623 CPU and an NVIDIA GTX 1080 Ti GPU. For each program tested, we run AFL-2.52b [76] on a single core machine for an hour to collect training data. The average number of training inputs collected for 9 programs is around 2K. We use 10KB as the threshold file size for selecting our training data from the AFL input corpus (on average 90% of the files generated by AFL were under the threshold).

| Fuzzer       | Technical Description            |
|--------------|----------------------------------|
| AFL [76]     | evolutionary search              |
| AFLFast [14] | evolutionary + markov-model-based search |
| FairFuzz [42] | evolutionary + byte masking      |
| Angora [20]  | evolutionary + dynamic-taint-guided + coordinate descent + type inference |
| Neuzz [64]   | Neural smoothing guided fuzzing   |

Table 1: State-of-the-art fuzzers used in our.

| Programs | # Lines | MTFuzz train (s) | Initial coverage |
|----------|---------|------------------|------------------|
| readelf-a | 21,647 | 703 | 3,132 |
| nm -C     | 53,457 | 202 | 3,031 |
| objdump -D | 72,955 | 703 | 3,939 |
| size      | 52,991 | 203 | 1,868 |
| strip     | 56,330 | 402 | 3,991 |
| harfbuzz-1.7.6 | 9,853  | 803 | 5,786 |
| libjpeg-9c | 8,857  | 1403 | 1,609 |
| mupdf-1.12.0 | 123,562 | 403 | 4,641 |
| libxml2-2.9.7 | 73,920 | 903 | 6,372 |
| zlib-1.2.11 | 1,893  | 107 | 1,438 |

Table 2: Test programs used in our study

RQ1: Performance

The first RQ assesses the performance of MTFuzz compared to 5 state-of-the-art fuzzers listed in Table 1 in terms of the number of real-world bugs detected (RQ1-A) and edge coverage (RQ1-B).
Table 3: Real-world bugs found after 24 hours by various fuzzers. MTFuzz finds the most number of bugs, i.e., 71 (11 unseen) comprised of 4 heap-oversflows, 3 Memory leaks, 2 integer overflows, and 2 out-of-memory bugs.

| Program | AFLFast | AFL | FairFuzz | Angora | Neuzz | MTFuzz |
|---------|---------|-----|----------|--------|-------|--------|
| readelf | 3       | 4   | 5        | 16     | 16    | 17     |
| nm      | 7       | 8   | 8        | 10     | 9     | 12     |
| objdump | 6       | 6   | 8        | 5      | 8     | 9      |
| size    | 4       | 4   | 5        | 7      | 6     | 10     |
| strip   | 5       | 7   | 9        | 20     | 20    | 21     |
| libjpeg | 0       | 0   | 0        | 0      | 0     | 1      |
| mupdf   | 0       | 0   | 0        | 0      | 0     | 1      |
| Total   | 27      | 29  | 35       | 58     | 60    | 71     |

Evaluation. To evaluate the number of bugs discovered by a fuzzer, we first instrument the program binaries with AddressSanitizer [1] and UnderdefinedBehaviorSanitizer [2]. Such instrumentation is necessary to detect bugs before crashes. Next, we run each of the fuzzers for 24 hours (all fuzzers use the same seed corpus) and gather the test inputs generated by each of the fuzzers. We run each of these test inputs on the instrumented binaries and count the number of bugs found in each setting. Finally, we use the stack trace of bug reports generated by two sanitizers to categorize the found bugs. Note, if multiple test inputs trigger the same bug, we only consider it once. Table 3 reports the results.

Observations. We find that:

(1) MTFuzz finds a total of 71 bugs, the most among other five fuzzers in 7 real world programs. In the remaining three programs, no bugs were detected by any fuzzer after 24 hours.

(2) Among these, 11 bugs were previously unreported.

Among the other fuzzers, Neuzz (another ML-based fuzzer) is the second best fuzzer, finding 60 bugs. Angora finds 58. We observe that the 11 new bugs predominantly belonged to 4 types: memory leak, heap overflow, integer overflow, and out-of-memory. Interestingly, MTFuzz discovered a potentially serious heap overflow vulnerability in mupdf that was not found by any other fuzzer so far (see Fig. 6). A mupdf function ensure_solid_xref allocates memory space for a single object (line 10) for each object of a pdf file and fills content to these memory chunks (line 14). Prior to that, at line 6, it tries to obtain the total number of objects by reading a field value xref->num_objects which is controlled by program input: a pdf

```
1 // mupdf-1.12.0-source/pdf/pdf-xref.c:174
2 static void ensure_solid_xref(...){
3 ...
4 int num = 1;
5 // xref->num_objects is manipulated by attacker
6 if (num < xref->num_objects) {
7     num = xref->num_objects;
8 ...
9 // allocate memory for num objects
10     new_sub->table = fz_calloc(ctx, num, sizeof(object));
11 ...
12 // fill content to num objects
13     for(i = 0; i < sub->len; i++) {
14         new_sub->table[i] = sub->table[i];
15     ...
```

Figure 6: Heap overflow bug in mupdf. The red line shows the bug.

Table 4: The average edge coverage of MTFuzz compared with other fuzzers after 24 hours runs for 5 repetitions.

| Programs    | MTFuzz | Neuzz | Angora | FairFuzz | AFL | AFLFast |
|-------------|--------|-------|--------|----------|-----|---------|
| readelf     | 6,701  | 6,759 | 6,514  | 3,423    | 1,072| 1,314   |
| nm          | 4,457  | 4,456 | 2,892  | 1,603    | 1,496| 1,270   |
| objdump     | 5,024  | 5,017 | 1,783  | 1,526    | 247 | 187     |
| size        | 3,728  | 3,737 | 2,107  | 1,954    | 1,426| 1,446   |
| strip       | 6,013  | 6,026 | 3,112  | 3,055    | 764 | 757     |
| libjpeg     | 1,189  | 1,199 | 499    | 977      | 671 | 850     |
| libxml      | 1,576  | 1,576 | 395    | 1,021    | 388 |         |
| mupdf       | 1,107  | 1,107 | 533    | 503      | 419 | 536     |
| zlib        | 298    | 297   | 196    | 294      | 229 |         |
| harfbuzz    | 6,325  | 6,329 | 2,616  | 5,615    | 3,692|         |

(a) Program binaries compiled with afl-gcc

(b) Program binaries compiled with afl-clang-fast

RQ1-B. How much edge coverage does MTFuzz achieve compared to other fuzzers?

Evaluation. To measure edge coverage, we run each of the fuzzers for 24 hours (all fuzzers use the same seed corpus). We periodically collect the edge coverage information of all the test inputs for each fuzzer using AFL’s coverage report toolkit afl1-showmap [76]. AFL provides coverage instrumentation scheme in two mainstream compilers GCC and Clang. While some authors prefer to use afl-gcc [14, 42, 64], some others use afl-clang-fast [20, 29]. The underlying compilers can have different program optimizations which affects how edge coverage is measured. Therefore, in order to offer a fair comparison with previous studies, we measure edge coverage on binaries compiled with both afl-gcc and afl-clang-fast. In the rest of the paper, we report results on programs compiled with afl-clang-fast. We observed similar findings with afl-gcc.
compared to both FairFuzz and AFLFast, MTFuzz covers significantly more edges, e.g. 4702 more than FairFuzz in `readelf` and over 28.1× more edges compared to AFLFast on `objdump`.

**Machine learning based fuzzers:** In comparison with the state of the art ML based fuzzer, NEUZZ [64], we observed that MTFuzz achieves much greater edge coverage in all 10 programs studied here. We notice improvements of 2000 more edges in `readelf` and 2500 more edges in `nm` and `strip`. Overall MTFuzz outperforms all the existing fuzzers on both `afl-gcc` (Tab. 4a) and `afl-c-clang` (Tab. 4b) binaries. In summary, MTFuzz found 71 real-world bugs (11 were previously unknown) and also reach on average 1,277 and up to 2,867 more edges compared to Neuzz, the second-best fuzzer, on 10 programs.

**RQ2: Contributions of Auxiliary Tasks**

MTFuzz is comprised of an underlying multi-task neural network (MTNN) that contains one primary task (edge coverage) and two auxiliary tasks namely, context-sensitive edge coverage and approach-sensitive edge coverage. A natural question that may arise is: How much does each auxiliary task contributes to the overall performance of MTFuzz?

**Evaluation.** To evaluate the contribution of each task, we investigate what would happen to the edge coverage when one of the auxiliary tasks is excluded from the multitask model. To answer this, we build the following four variants of the underlying MTNN:

1. **(EC):** A single-task NN with only the primary task of predicting edge coverage.
2. **(EC, Call Ctx):** An MTNN with edge coverage as the primary task and context-sensitive edge coverage as the auxiliary task.
3. **(EC, Approach):** An MTNN with edge coverage as the primary task and approach-sensitive edge coverage as the auxiliary task.
4. **MTFuzz:** Our proposed model with edge coverage as the primary task and two auxiliary tasks context-sensitive edge coverage and approach-sensitive edge coverage.

We rule out other possible confounders by ensuring that: (a) each setting shares the same hyper-parameters; (b) all subsequent steps in fuzzing remain the same across each experiment; and (c) each variant of the multi-task model is given the same initial seed corpus. With the above settings, we run each of the above multi-task models on all our programs from Tab. 2 for 1 hour to record the edge coverage for each of these MTNN models.

**Observations.** Our results are tabulated in Tab. 5. We make the following noteworthy observations:

1. Fuzzer that uses an MTNN trained on edge coverage as the primary task and context-sensitive edge coverage as the only auxiliary task tends to perform only marginally better than a single task NN based on edge coverage. In some cases, e.g., in Tab. 5 we notice about 25% more edges. However, in some other cases, for example in `libjpeg`, we noticed that the coverage reduces by almost 31%.
2. The above trend is also observable for using edge coverage with approach-sensitive edge coverage as the auxiliary. For example, in `libjpeg`, the edge coverage is lower than the single-task model that uses only edge coverage.
3. However, MTFuzz, which uses both context-sensitive edge coverage and approach-sensitive edge coverage as auxiliary tasks to edge coverage, performs noticeably better than all other models with up to **800 more edges covered** (≈20%) in the case of `readelf`.

The above behavior is expected because each auxiliary task provides very specific albeit somewhat partial context to edge coverage —context-sensitive edge coverage only provides context to triggered edges, while approach-sensitive edge coverage only reasons about non-triggered edges (see §3.1 for details). Used in isolation, a partial context does not have much to offer. However, while working together as auxiliary tasks along with the primary task, it provides a better context to edge coverage resulting in overall increased edge coverage (see the last column of Tab. 5).
MTFuzz benefits from both the auxiliary tasks, i.e., context-sensitive edge coverage and edge coverage, used along with the primary task—predicting edge coverage with up to **20%** more edge coverage.

**RQ3. Impact of Design Choices**

While building MTFuzz, we made few key design choices such as using a task-specific adaptive loss (§3.2.2) to improve the quality of the multi-task neural network (MTNN) model and a novel seed selection strategy based on importance sampling (see §3.4). Here we assess how helpful these design choices are.

**RQ3-A. What are the benefits of using adaptive loss?**

MTNN model predicting for edge coverage and for context-sensitive edge coverage tend to experience severely imbalanced class labels. Consider the instance when a certain input triggers an edge for the first time. This is an input of much interest because it represents a new behaviour. The MTNN model must learn what leads to this behaviour. However, in the training sample, there exists only one positive sample for this new edge in the entire corpus. An MTNN that is trained with an off-the-shelf loss function is likely to misclassify these edges resulting in a false negative error. Such false negatives are particularly damaging because a large number of new edge discoveries go undetected affecting the overall model performance. To counter this, we defined an adaptive loss in §3.2.2; here we measure how much it improves the MTNN’s performance.

**Evaluation.** To evaluate the effect of class imbalance, we measure recall which is high when the overall false negatives (FN) are low. While attempting to minimize FNs the model must not make too many false positive (FP) errors. Although false positives are not as damaging as false negatives, we must attempt to keep them low. We therefore also keep track of the F1-scores which quantify the trade-off between false positives and false negatives. We train MTFuzz with two different losses (i.e., with our adaptive loss and with the default cross-entropy loss) on 10 programs for 100 epochs and record the final recall and F-1 scores.

**Observations.** The result is shown in Tab. 6. We observe that adaptive loss results in MTNNs with an average of 90% recall score on 10 programs, while the default loss model only achieves on average 75% recall score. Generally, we notice improvements greater than **15%** over default loss functions. The low recall for default loss function indicates that it is susceptible to making a lot of false negative predictions. However, our adaptive loss function is much better at reducing false negative predictions. Also, the adaptive loss model achieves on average F-1 score of 72%, while unweighted loss model achieves an average of 70%. This is encouraging because even after significantly reducing the number of false negatives, we maintain the overall performance of the MTNN.

Importance sampling helps MTFuzz achieve on average 1.66× edge coverage compared with random seed selection on all 10 programs.

| Programs | Adaptive Recall(%) | F1(%) | Default Recall(%) | F1(%) | Seed Selection |
|----------|-------------------|-------|-------------------|-------|----------------|
| readelf  | 88                | 68    | 74                | 66    | 4,799          |
| nm       | 89                | 62    | 69                | 62    | 577            |
| objdump  | 89                | 72    | 65                | 71    | 672            |
| size     | 94                | 81    | 78                | 78    | 502            |
| strip    | 89                | 73    | 80                | 72    | 954            |
| harfbuzz | 92                | 67    | 80                | 71    | 884            |
| libjpeg  | 88                | 66    | 65                | 65    | 223            |
| mupdf    | 92                | 84    | 90                | 84    | 269            |
| libxml2  | 90                | 70    | 76                | 69    | 699            |
| zlib     | 86                | 70    | 70                | 65    | 677            |

**RQ3-B. How does seed-selection help?**

**Evaluation.** Here, we evaluate our seed selection strategy (§3.4) by comparing it to a random selection strategy. Specifically, we run two variants of MTFuzz, one with importance sampling for seed selection and the other with a random seed selection. All other components of the tool such as MTNN model, hyperparameters, random seed, etc. are kept constant. We measure the edge coverage obtained by both the strategies on 10 programs after fuzzing for one hour. Tab. 6 shows the results.

**Observations.** When compared to a random seed selection strategy, importance sampling outperforms random seed selection in all 10 programs offering average improvements of 1.66× more edges covered than random seed selection—for readelf, it covers around 2000 more edges. This makes intuitive sense because, the goal of importance sampling was to retain the newly generated inputs that invoke certain rare edges. By populating the corpus with such rare and novel inputs, the number of newly explored edges would increase over time, resulting in increase edge coverage (see Tab. 6).

RQ4. Transferability

In this section, we explore the extent to which MTFuzz can be generalized across different programs operating on the same inputs (e.g., two ELF fuzzers). Among such programs, we study if we can transfer inputs generated by fuzzing from one program to trigger edge coverage in another program (RQ4-A) and if it is possible to transfer the shared embedding layers between programs (RQ4-B).

**RQ4-A. Can inputs generated for one program be transferred to other programs operating on the same domain?**

MTFuzz mutates the hot-bytes in the inputs to generate additional test inputs. These hot-bytes are specific to the underlying structure of the inputs. Therefore, inputs that have been mutated on these hot-bytes should be able to elicit new edge coverage for any program that parses the same input.

**Evaluation.** To answer this question, we explore 5 different programs that operate on 2 file types: (1) readelf, size, and nm operating on ELF files, and (2) libxml and xmlwf [3] operating on XML files. For all the programs that operate on the same file format:

1. We pick a source program (say $S = P_i$) and use MTFuzz to fuzz the source program for 1 hour to generate new tests inputs.
2. Next, for every other target program $T = P_j$, we use the test inputs from the previous step to measure the coverage. Note that we do not deploy the fuzzer on the target program we merely measure the code coverage.


5 Threats to Validity

(a) Initialization: For the fuzzers studied here, it is required to provide initial set of seed inputs. To ensure a fair comparison, we use the same set of seed inputs for all the fuzzers.
(b) Target programs: We selected diverse target programs from a wide variety of software systems. One still has to be careful when generalizing to other programs not studied here. We ensure that all the target programs used in this study have been used previously; we do not claim that our results generalize beyond these programs.
(c) Other fuzzers: When comparing MTFuzz with other state-of-the-art fuzzers, we use those fuzzers that are reported to work on the programs tested here. Our baseline fuzzer NEUZZ [64] has reported to outperform many other fuzzers on the same studied programs. Since we are outperforming NEUZZ, it is reasonable to expect that we will outperform the other fuzzers as well.

6 Related Work

Fuzzing [51] has garnered significant attention recently. There are three broad types of fuzzers: (a) Blackbox [19, 36, 37] with no knowledge of the target program, (b) Whitebox [16, 31, 32, 62] with source/binary level access the target program, and (c) Greybox fuzzers like AFL with the ability to instrument and collect some target-program-specific information like code coverage. This paper specifically focuses on greybox fuzzers. Most greybox fuzzers use evolutionary search to guide their input generation strategy [76]. Since the release of AFL [76], the researchers have attempted to implement a wide range of mutation strategies augmented with program analysis to optimize the evolutionary mutation process [8, 13, 14, 20, 42, 42, 56, 74, 75]. All of these projects focus on manually designing different mutation strategies and use either program analysis [8, 13, 20] or aggregate statistics [42] to customize their strategy for specific target programs. By contrast, MTFuzz uses multi-task neural networks to automatically learn a compact representation of input to identify and mutate the hot-bytes.

More recently, machine learning techniques are being increasingly used to improve fuzzing. One line of work focused on using neural networks to model the input format of the target programs based on a corpus of sample inputs [9, 15, 33, 59, 64]. Another alternative approach like NEUZZ [64] models the edge behaviour of a program using a neural network. In this paper, we demonstrate that neural networks can be further used to adaptively learn a number of mutation parameters that can significantly improve edge coverage.

Transfer learning [28, 35, 50, 57, 65] is beneficial when there is insufficient data for a target task but there exists sufficient data for other source task. To ensure ideal transfer, the target task and source task should have same or similar feature space and share similar data distribution. Rana et al. [58] and Dai et al. [24] [25] use transfer learning to perform cross-domain text classification. Long et al. [45] and Sun and Saenko [68] apply transfer learning to solve image-classification problem. We demonstrate that MTFuzz can transfer a NN learnt on one program to other similar programs.

7 Conclusion

This paper presents MTFuzz, a multi-task neural-network fuzzing framework. MTFuzz learns from multiple code coverage measures to reduce a sparse and a high-dimensional input space to a compact representation of the input, it should, in theory, allow for these compact representations to be transferred across programs that share the same input, e.g., across programs that process ELF binaries.

MTFuzz’s compact embedding can be transferred across programs that operate on similar input formats. We achieve up to 4x more edge coverage for XML files (with an average of 2x more edge coverage across all programs) compared to other fuzzers.
representation. This compact representation is used to guide the fuzzier towards unexplored regions of the source code. Further, this compact representation can be transferred across programs that operate on the same input format. Our findings suggest MTUffz can improve edge coverage significantly while discovering several previously unseen bugs.

REFERENCES

[1] 2020. Address Sanitizer, Thread Sanitizer, and Memory Sanitizer. https://github.com/google/sanitizers.

[2] 2020. UndefinedBehaviorSanitizer. https://clang.llvm.org/docs/UndefinedBehaviorSanitizer.html.

[3] 2020. xmlbf xml parser. https://libexpat.github.io/.

[4] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A system for large-scale machine learning. In 12th USENIX Symp. on Operating Systems Design and Implementation (OSDI) 265–283.

[5] Andrea Arcuri. 2010. It does matter how you normalise the branch distance in search based software testing. In 3rd Intl Conf. Software Testing, Verification and Validation. Iee, 205–214.

[6] Andreas Argyrinou, Theodoros Evgeniou, and Massimiliano Pontil. 2008. Convex multi-task feature learning. Machine learning 73, 3 (2008), 243–272.

[7] Abhishek Arya and Cris Necker. 2012. Fuzzing for security. 1 (2012), 2013. https://blog.chromium.org/2012/04/fuzzing-for-security.html

[8] Marcelus Aschermann, Robert Gawkil, and Thorsten Holz. 2019. REDQUEEN: Fuzzing with Input-to-State Correspondence. In Proceedings of the Network and Distributed System Security Symposium (2019).

[9] Osbert Bastani, Rahul Sharma, Alex Aiken, and Percy Liang. 2017. Synthesizing Thorsten Holz. 2019. REDQUEEN: Fuzzing with Input-to-State Correspondence. In Proceedings of the Network and Distributed System Security Symposium (2019).

[10] Anthony Baxter. 1995. A Bayesian/information theoretic model of learning to learn via multiple task sampling. Machine learning 28, 1 (1997), 7–39.

[11] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35, 8 (2013), 1798–1828.

[12] Joachim Bünge and Andreas Sogaard. 2017. Identifying benefit relations for multi-task learning in deep neural networks. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference 2, 2016 (2017), 164–169. https://doi.org/10.18653/v1/e17-2016.arXiv:1702.08303

[13] Tim Blazytko, Cornelius Aschermann, Moritz Schlägel, Ali Abbasi, and Perey Liang. 2017. Synthesizing Program Input Grammars. SKPLAN Not. 52, 6 (June 2017), 95–110. https://arxiv.org/abs/1114.0587v3

[14] Jonathan Baxter. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. Machine learning 28, 1 (1997), 7–39.

[15] Konstantin Bottinger, Patrice Godefroid, and Rishabh Singh. 2018. Deep reinforcement fuzzing. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 761–770.

[16] Matthias Hörschele and Andreas Zeller. 2016. Mining Input Grammars from Dynamic Taints. In Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering (IEEE/ACM), New York, NY, USA, 620–215. https://doi.org/10.1145/2774230.2864586

[17] Bo Jiang, Ye Liu, and WK Chan. 2018. ContractFuzzer: Fuzzing smart contracts for vulnerability detection. In Proceedings of the 33rd IEEE/ACM Conference on Automated Software Engineering. ACM, 259–269.

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 2017. Variational inference of disentangled latent concepts from unlabeled observations. arXiv preprint arXiv:1711.00848 (2017).

[19] Caroline Lemieux and Koushik Sen. 2018. FairFuzz: Targeting rare branches to rapidly increase greybox fuzz testing coverage. In Proceedings of the 33rd ACM SIGSOFT International Conference on Automated Software Engineering. ACM.

[20] Yuekang Li, Bihuan Chen, Mahinthan Chandramohan, Shang-Wei Lin, Yang Liu, and Alwen Tiu. 2017. Steelix: Program-State Based Binary Fuzzing. In Proceedings of the 11th Joint Meeting on Foundations of Software Engineering (FSE). ACM.

[21] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2018. Deep convolutional inverse graphics network. In Advances in neural information processing systems. 2539–2547.

[22] Abhishek Kumar, Prasanna Sattigeri, and Inalavishal Balakrishnan. 2017. Variational inference of disentangled latent concepts from unlabeled observations. arXiv preprint arXiv:1711.00848 (2017).

[23] Pinghua Gong, Jiayu Zhou, Wei Fan, and Jieping Ye. 2014. Efficient multi-task feature learning with calibration. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 761–770.

[24] Shuitao Gan, Chao Zhang, Peng Chen, Bodong Zhao, Xiaojun Qin, Dong Wu, and Zuoqin Chen. 2020. GREYONE: Data Flow Sensitive Fuzzing. In 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity20/presentation/gan

[25] Patrice Godefroid, Adam Kiezun, and Michael Y. Levin. 2018. Grammar-based Whitebox Fuzzing. In Proc. 29th ACM SIGPLAN Conf. Programming Language Design and Implementation (Tucson, AZ, USA) (PAtl 08). ACM, New York, NY, USA, 206–215. https://doi.org/10.1145/1375581.1375607

[26] Patrice Godefroid, Nils Klarlund, and Koushik Sen. 2005. DART: directed automated random testing. In ACM Sigplan Notices, Vol. 40. ACM, 213–223.

[27] Patrice Godefroid, Michael Y. Levin, and David A Molnar. 2008. Automated Whitebox Fuzz Testing. In Proceedings of the Network and Distributed System Security Symposium 8 (2008), 151–166.

[28] Patrice Godefroid, Hila Peleg, and Rishabh Singh. 2017. Learn&fuzz: Machine learning for input fuzzing. In Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering. ACM, 50–59.

[29] Phil McMinn and Mike Holcombe. 2004. Hybridizing Evolutionary Testing and multi-task feature learning. In Proceedings of the 3rd IEEE/ACM International Conference on Automated Software Engineering. ACM.

[30] girlon, Gaurav Jaiswal, and Masoud Javidi. 2016. Transferable Features with Deep Adaptation Networks. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (Lille, France) (ICML’15).

[31] Varun Girdhar, Sergey Brin, and Adam Tzeng. 2014. Deep Multi-task Automatic Triage and Classification. In Proceedings of the 31st Annual International Conference on Machine Learning (ICML-14). JMLR.org.

[32] Varun Girdhar, Sergey Brin, and Adam Tzeng. 2014. Deep Multi-task Automatic Triage and Classification. In Proceedings of the 31st Annual International Conference on Machine Learning (ICML-14). JMLR.org.

[33] Varun Girdhar, Sergey Brin, and Adam Tzeng. 2014. Deep Multi-task Automatic Triage and Classification. In Proceedings of the 31st Annual International Conference on Machine Learning (ICML-14). JMLR.org.

[34] Varun Girdhar, Sergey Brin, and Adam Tzeng. 2014. Deep Multi-task Automatic Triage and Classification. In Proceedings of the 31st Annual International Conference on Machine Learning (ICML-14). JMLR.org.
