Fuzzing Deep-Learning Libraries via Automated Relational API Inference

Yinlin Deng∗
University of Illinois at Urbana-Champaign
yinlind2@illinois.edu

Chenyuan Yang∗
University of Illinois at Urbana-Champaign
cy54@illinois.edu

Anjiang Wei
Stanford University
anjiang@stanford.edu

Lingming Zhang
University of Illinois at Urbana-Champaign
lingming@illinois.edu

ABSTRACT
Deep Learning (DL) has gained wide attention in recent years. Meanwhile, bugs in DL systems can lead to serious consequences, and may even threaten human lives. As a result, a growing body of research has been dedicated to DL model testing. However, there is still limited work on testing DL libraries, e.g., PyTorch and TensorFlow, which serve as the foundations for building, training, and running DL models. Prior work on fuzzing DL libraries can only generate tests for APIs which have been invoked by documentation examples, developer tests, or DL models, leaving a large number of APIs untested. In this paper, we propose DeepREL, the first approach to automatically inferring relational APIs for more effective DL library fuzzing. Our basic hypothesis is that for a DL library under test, there may exist a number of APIs sharing similar input parameters and outputs; in this way, we can easily “borrow” test inputs from invoked APIs to test other relational APIs. Furthermore, we formalize the notion of value equivalence and status equivalence for relational APIs to serve as the oracle for effective bug finding. We have implemented DeepREL as a fully automated end-to-end relational API inference and fuzzing technique for DL libraries, which 1) automatically infers potential API relations based on API syntactic/semantic information, 2) synthesizes concrete test programs for invoking relational APIs, 3) validates the inferred relational APIs via representative test inputs, and finally 4) performs fuzzing on the verified relational APIs to find potential inconsistencies. Our evaluation on two of the most popular DL libraries, PyTorch and TensorFlow, demonstrates that DeepREL can cover 157% more APIs than state-of-the-art FreeFuzz. To date, DeepREL has detected 162 bugs in total, with 106 already confirmed by the developers as previously unknown bugs. Surprisingly, DeepREL has detected 13.5% of the high-priority bugs for the entire PyTorch issue-tracking system in a three-month period. Also, besides the 162 code bugs, we have also detected 14 documentation bugs (all confirmed).

CCS CONCEPTS
• Software and its engineering → Software testing and debugging; Software reliability.

KEYWORDS
Fuzz Testing, Oracle Inference, Deep Learning, Differential Testing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ESEC/FSE ’22, November 14–18, 2022, Singapore, Singapore
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9413-0/22/11...$15.00
https://doi.org/10.1145/3540250.3549085

1 INTRODUCTION
Recent years have witnessed the surge of deep learning (DL) in a variety of applications, including computer vision [32, 60], natural language processing [24, 28], robotics [23, 35], bioinformatics [46, 57], and software engineering [25, 29, 36, 37, 45, 46, 78, 79, 81, 82]. Meanwhile, similar as traditional software systems, DL systems can also have bugs, which can lead to serious consequences and may even threaten human lives [4].

To date, most prior work on DL testing focused on testing/verifying DL models, with an emphasis on adversarial attacks [14, 19, 27, 42, 47, 51], metrics for model testing [31, 34, 41, 53, 77], application-specific model testing [62, 85, 88], and verifying certain properties of models [15, 38]. Meanwhile, there is limited work targeting the reliability of DL libraries, which serve as the central infrastructures for building DL models, and are the foundation for training and deploying DL models. CRADLE [54] is one of the trailblazing work on testing DL libraries, which resolves the oracle challenge with differential testing of various DL models on multiple backends of Keras [1]. AUDEE [30] and LEMON [65] further augment CRADLE by leveraging search-based mutation strategies to generate more diverse DL models/inputs for testing library code. Different from the above model-level DL library testing techniques, more recently, FreeFuzz [66] has been proposed to mine example inputs from open source (including code snippets from the library documentation, developer tests, and DL models in the wild) to directly test each DL library API in isolation. FreeFuzz has been evaluated on PyTorch [52] and TensorFlow [13], currently the two most popular DL libraries (with 54K/162K stars on Github). The experimental results show that FreeFuzz can cover 9x more APIs than state-of-the-art LEMON and detect various previously unknown bugs.

∗Both authors contributed equally to this research.
Despite the promising results, existing techniques for fuzzing DL libraries still suffer from the following limitations. First, the input generation is still far from optimal. CRADLE and AUDEE can only test APIs that are covered in the original models, and LEMON can cover slightly more APIs with layer mutations; furthermore, although FreeFuzz can cover up to 1158 APIs for PyTorch and TensorFlow (which is already a huge improvement over other work), it is still unable to test an API if there is no code snippet directly invoking the API. Second, there is still a lack of powerful test oracles. Existing techniques typically perform differential testing across different DL libraries or hardware backends (e.g., GPU/CPU) to address the test oracle issue. However, differential testing across DL libraries is typically applied at the model level and suffers from the limited effectiveness of model-level testing (e.g., limited API coverage and accumulated floating-point precision loss) [30, 54, 65], while different backends often share common code logic/design (and thus may also share similar bugs) [66]. Thus, it is also crucial to investigate novel test oracles for effective DL library fuzzing.

To address the aforementioned limitations, in this work, we open a new dimension for testing DL libraries via automated relational API inference. The inspiration stems from the fact that prior work [16, 26, 43, 64] has discovered a number of equivalent APIs in traditional software systems (e.g., Java projects). We envision such relational API inference also to be an inspiring direction for fuzzing DL libraries. In this way, given the same inputs generated via fuzzing, APIs that are equivalent in functionality should produce the same numerical results (i.e., value equivalence). Moreover, besides the previously studied equivalent APIs, we further leverage the fact that DL APIs with similar functionality should behave similarly in terms of program status (i.e., status equivalence) for more effective fuzzing. For example, although AdaptiveAvgPool3d and MaxPool3d in PyTorch are not equivalent, they are functionally similar APIs; thus, we can feed any valid input of the first API to the second API and expect its invocation to also be successful. Based on this intuition, we can easily “borrow” test inputs generated for one API to test other relational APIs. Also, API relations can directly serve as test oracles for differential testing. Therefore, we can easily overcome the aforementioned limitations.

We have built a fully-automated technique, DeepREL, which infers such API relations without human intervention for fuzzing DL libraries. One key challenge is how to obtain the API relations automatically and accurately. Existing work [16, 26, 43, 64] on equivalent API inference for traditional software systems can hardly be applied for DL library testing, e.g., the most recent MeMo work [16] heavily relies on well-documented API relations, which are rare in DL libraries. To this end, DeepREL first automatically infers all possible candidate API pairs that are matched based on API syntactic and semantic information. Then, DeepREL synthesizes concrete test programs for those potentially relational APIs. After that, DeepREL leverages a set of representative valid inputs (automatically traced during prior normal API executions) to check whether the inferred API relations hold or not. Lastly, DeepREL takes the validated API pairs and leverages mutation-based fuzzing to generate a much more diverse and extensive set of test inputs for detecting potential

1Some of such existing work (16, 26, 43) treated the entire software systems under test as the test objects, and thus viewed this as metamorphic testing [20]. In this paper, we treat each API as a test object and view this as differential testing [44].
computing the loss (\(\text{loss} = \text{CrossEntropyLoss(label, out)}\)), computing the gradient (\(\text{loss.backward()}\)), and invoking the optimizer (\(\text{optimizer.step()}\)) for back-propagation.

**DL APIs.** When running a DL model, the APIs involved in building the model are also executed. Essentially, writing a DL program can be viewed as defining a *computation graph*. It is a directed acyclic graph (DAG) whose nodes are DL APIs while edges are the flow of the tensors. Figure 1 also shows the computation graph, composed of three parts: forward part (taking input tensors and weight tensors as input), loss computation part (requiring label tensors), and backward part (for updating weight tensors). Actually, the backward part needs to construct a very complex graph [87], but we omit it in Figure 1 for simplicity. Essentially, running a whole DL model can be broken down into invoking a series of DL APIs based on the topological sorting [74] of the computation graph.

### 2.2 Fuzzing DL Libraries

To our knowledge, there are mainly two categories of work for fuzzing DL libraries, model-level testing, and API-level testing. **Model-level Testing.** CRADLE [54] is one of the first to apply differential testing for DL libraries. Since Keras [1] is a library featuring high-level APIs for building DL models, APIs in Keras may have multiple implementations in its supported lower-level libraries. Thus, CRADLE takes 30 pre-trained DL models as input and runs differential testing to find inconsistencies between different low-level libraries for Keras. More recently, AUDEE [30] and LEMON [65] have proposed to use search-based mutation strategies to generate mutated DL models for differential testing on different backends. While AUDEE focuses on mutating parameters of layers, weight tensors, and input tensors, LEMON applies mutation rules by adding/deleting layers and changing the values of weight tensors. In this way, LEMON’s mutation rules are more general and can cover more APIs than the original models. However, even LEMON’s mutation can only be applied to a limited number of APIs. For instance, its intact-layer mutation rule [65] requires that the output tensor shape of the API to be added/deleted should be identical to its input tensor shape, making a large number of APIs inapplicable for mutation. Researchers have recently shown that LEMON can hardly invoke additional library code/APIs with its mutation rules, e.g., covering only 35 APIs in total for TensorFlow [66]. Very recently, Liu et al. [39] proposed NNSmith, which leverages lightweight specifications for more diverse model generation. Meanwhile, it also mainly focuses on commonly used APIs/operators.

**API-level Testing.** Different from prior work on DL library fuzzing, the recent FreeFuzz work [66] proposes to directly mine test inputs from open source for API-level fuzzing, a much finer-grained level than model-level testing. One challenge is that Python is a dynamically-typed language, making it hard to determine the types of API parameters for fuzzing Python APIs. Prior work has to manually set up the API arguments, and thus can only test a small number of APIs, e.g., Predo [86] can only test 7 APIs for TensorFlow. A very recent work DocTer [76] constructs rules to extract DL-specific input constraints from API documentation and uses them to generate valid/invalid inputs for testing DL libraries; meanwhile, it requires manual annotation for 30% of API parameters. In contrast, FreeFuzz resolves this challenge fully automatically via dynamically tracing API executions in code snippets from the documentation, developer tests, and 202 DL models. More specifically, FreeFuzz records the traced argument values in a database and further performs mutation-based fuzzing to mutate those traced values to generate even more inputs for fuzzing DL library APIs. Lastly, FreeFuzz applies differential testing on different hardware backends (i.e., CPU/GPU) for detecting potential consistency bugs. Despite its big improvement over prior work, FreeFuzz can only test 1158 APIs for PyTorch and TensorFlow, which are the ones covered in its input mining stage, leaving a total of 6815 APIs uncovered. Also, different hardware backends may still share code logic/design, causing the differential testing oracle used by FreeFuzz to miss various bugs. In this work, we propose to test relational APIs to further overcome such limitations. We build our technique (DeepREL) upon FreeFuzz to automatically infer relational APIs and leverage them to fuzz DL libraries. Note, however, that our DeepREL idea is general and can be built on any API-level fuzzer for DL libraries (e.g., DocTer [76]). We choose FreeFuzz since it is a recent state-of-the-art technique that is both publicly available and fully automated.

### 3 PRELIMINARIES

We first introduce the preliminaries for our fuzzing technique in this section. Given the set of all possible APIs, \(\mathcal{A}\), for a DL library under test, we aim to define the relational property between the invocation results of a source API \(S \in \mathcal{A}\) and a target API \(T \in \mathcal{A}\).

Intuitively, we can directly check whether \(S\) and \(T\) produce equivalent outputs. For example, Figure 2 shows an API pair which, according to the PyTorch documentation [5], should always produce the same results. The `broadcast_shapes` API applies broadcasting on a list of compatible shapes to align them, while the `broadcast_tensors` API applies broadcasting on a list of shape-compatible tensors to align their shapes. In fact, the first API can be rewritten as 1) creating intermediate empty tensors from tensor shapes with `map` and `torch.empty`, 2) calling `broadcast_tensors` with these tensors, and 3) getting the shape of the output tensor. Since the source and target APIs can achieve the same functionality with totally different implementations given the same input (shapes in Figure 2), they provide a great opportunity for differential testing [44]. Therefore, we have the following formal definition:

**Definition 3.1.** *Equivalence* \(\text{Equivalence}_{\text{value}}\). Given a set of inputs \(\mathcal{D}\), source API \(S \in \mathcal{A}\) and target API \(T \in \mathcal{A}\) satisfy *Equivalence* \(\text{value} \) (modulo \(\mathcal{D}\)) if their invocations always output the same results given any input in \(\mathcal{D}\). Formally,

\[
S \equiv T(\text{mod } \mathcal{D}) \iff \forall x \in \mathcal{D}. S(x) = T(x) 
\]  

(1)

While this can be effective in detecting potential consistency bugs, the checking is too strict and may not apply to a large number
of APIs. In fact, it could be possible that $S$ and $T$ produce totally different results, but tend to behave similarly when given similar inputs. For example, the API pair shown in Figure 3 does not satisfy the Equivalence-value property since the output of AdaptiveAvgPool3d is different from AdaptiveMaxPool3d. The first API applies a 3D adaptive average pooling over an input but the latter applies a 3D adaptive maximum pooling. However, these two APIs do share certain functionality in common since both of them apply a pooling operation, which is also valuable for testing. Therefore, we further abstract the invocation result of a program into a set of coarse-grained statuses: Success, Exception, and Crash. Success denotes that program executions terminate normally, while Exception means that program executions throw known exceptions. Lastly, Crash represents the cases where the program executions crash with unexpected errors, e.g., segmentation faults or INTERNAL ASSERT FAILED errors (which are “never acceptable” as commented by PyTorch developers). We then further introduce the notation of $\langle \cdot \rangle \in \mathcal{A}$ (Success, Exception, Crash) to return the execution status of an input program. For example, $\langle S(x) \rangle = \text{Success}$ indicates that $S$ terminates normally with the input $x$. In this way, we can define another property for checking potential consistency:

**Definition 3.2. Equivalence-status.** Given a set of inputs $\mathcal{D}$, source API $S \in \mathcal{A}$ and target API $T \in \mathcal{A}$ satisfy Equivalence-status (modulo $\mathcal{D}$) iff their invocation always output the same statuses given any input in $\mathcal{D}$. Formally,

$$S \sim T(\text{mod } \mathcal{D}) \iff \forall x \in \mathcal{D}. \langle S(x) \rangle = \langle T(x) \rangle \quad (2)$$

To conclude, the Equivalence-status relation is a relaxed notation for the Equivalence-value relation, which is in turn a relaxed notion of semantic equivalence (denoted as $S \equiv T$). Formally,

$$S \equiv T \Rightarrow S \equiv T(\text{mod } \mathcal{D}) \Rightarrow S \sim T(\text{mod } \mathcal{D}) \quad (3)$$

One crucial component of these two definitions is the domain $\mathcal{D}$ on which the properties are constrained. Aiming for more accurate API relations, it would be beneficial to cover more representative APIs. In fact, it could be possible that $S$ and $T$ produce totally different results, but tend to behave similarly when given similar inputs. For example, the API pair shown in Figure 3 does not satisfy the Equivalence-value property since the output of AdaptiveAvgPool3d is different from AdaptiveMaxPool3d. The first API applies a 3D adaptive average pooling over an input but the latter applies a 3D adaptive maximum pooling. However, these two APIs do share certain functionality in common since both of them apply a pooling operation, which is also valuable for testing. Therefore, we further abstract the invocation result of a program into a set of coarse-grained statuses: Success, Exception, and Crash. Success denotes that program executions terminate normally, while Exception means that program executions throw known exceptions. Lastly, Crash represents the cases where the program executions crash with unexpected errors, e.g., segmentation faults or INTERNAL ASSERT FAILED errors (which are “never acceptable” as commented by PyTorch developers). We then further introduce the notation of $\langle \cdot \rangle \in \mathcal{A}$ (Success, Exception, Crash) to return the execution status of an input program. For example, $\langle S(x) \rangle = \text{Success}$ indicates that $S$ terminates normally with the input $x$. In this way, we can define another property for checking potential consistency:

**Definition 3.2. Equivalence-status.** Given a set of inputs $\mathcal{D}$, source API $S \in \mathcal{A}$ and target API $T \in \mathcal{A}$ satisfy Equivalence-status (modulo $\mathcal{D}$) iff their invocation always output the same statuses given any input in $\mathcal{D}$. Formally,

$$S \sim T(\text{mod } \mathcal{D}) \iff \forall x \in \mathcal{D}. \langle S(x) \rangle = \langle T(x) \rangle \quad (2)$$

To conclude, the Equivalence-status relation is a relaxed notation for the Equivalence-value relation, which is in turn a relaxed notion of semantic equivalence (denoted as $S \equiv T$). Formally,

$$S \equiv T \Rightarrow S \equiv T(\text{mod } \mathcal{D}) \Rightarrow S \sim T(\text{mod } \mathcal{D}) \quad (3)$$

One crucial component of these two definitions is the domain $\mathcal{D}$ on which the properties are constrained. Aiming for more accurate API relations, it would be beneficial to cover more representative test inputs within the intersection of the valid input space of the source and target APIs as the domain.

### 4 FUZZING RELATIONAL APIS

Figure 4 shows the overview of our DeepREL technique for fuzzing relational APIs of DL libraries. DeepREL takes as input the targeted DL library, its API documentation, and a database of valid historical API invocations (e.g., automatically collected via running documentation examples, library tests, and DL models [66]). Each entry of the database contains the concrete argument values passed into an API during an invocation and is obtained through dynamic tracing. Overall, DeepREL performs the following four phases iteratively:

**API Matcher (Section 4.1).** In order to test a DL library which typically has hundreds or even thousands of APIs, the first challenge is to identify the API pairs that are likely to satisfy the desired properties Equivalence-value or Equivalence-status. API Matcher maps each API into embeddings based on API documentation, and uses embedding similarity to identify candidates of API pairs.

**Invocation Synthesizer (Section 4.2).** Given a collection of potential API pairs, Invocation Synthesizer decides how to invoke them. To construct valid invitations for later verification, we impose a constraint on the source API: it must have at least one (valid) invocation in the database. In this way, given an invocation of the source API, Invocation Synthesizer aims to synthesize the invocation code for the target API.

**API Match Verifier (Section 4.3).** Given the invocation code of matched APIs, this phase would check whether each API pair satisfies property Equivalence-value or Equivalence-status with a set of representative inputs as the verifying test inputs. If the result values (resp. execution statuses) are consistent for all tests, then API Match Verifier accepts the API pair as Equivalence-value (resp. Equivalence-status). If API Match Verifier detects any inconsistency in this phase, then it rejects the API pair.

**API Fuzzer (Section 4.4).** The last step is to leverage the verified API pairs to detect potential consistency bugs. API Fuzzer uses mutation-based fuzzing to generate a large number of test inputs for source APIs, and tests the verified API pairs with oracles (Section 3).

Lastly, recall that in order to generate valid inputs, the source API must have at least one (valid) invocation in the database. DeepREL further adopts an iterative process to cover more API pairs (Section 4.5). The newly generated valid target API invocations can be added to the database to serve as the source APIs for the next iterations to detect more potential API pairs. The following sections would explain each phase in detail.

#### 4.1 API Matcher

In this phase, DeepREL identifies potential matched API pairs from the documentation. DeepREL uses API Matcher to infer similar API pairs as candidate pairs (which will be further verified later). API Matcher would map each API into its embeddings, and compute the cosine similarity of embeddings to be the similarity of each API pair. We consider signature similarity and document similarity that cover both API syntactic and semantic information for similarity computation. Overall, for each API pair, the similarity is defined as the maximum of the two:

$$\text{Sim}_{API}(S, T) = \text{Max}(\text{Sim}_{sig}(S, T), \text{Sim}_{doc}(S, T)) \quad (4)$$

We compute the pair-wise similarity for every API pair, and pair each API with its $K$-closest neighbors as the candidate pairs. $K$ is a hyper-parameter and is set to 10 in the default setting of DeepREL. Notably, we also analyze the impact of different values of $K$ in our experimental study (Section 6.3).

**Signature Similarity.** The signature of an API contains the API name and an ordered list of argument names. The APIs that will be paired typically have signatures that follow a similar syntactic pattern. For example, the API pair $\text{tf.maximum}$ and $\text{tf.minimum}$ satisfies Equivalence-status, and their signatures are very similar: $\text{tf.maximum}(x, y)$ and $\text{tf.minimum}(x, y)$. We map an API signature into its TF-IDF (term frequency-inverse document frequency) embedding [73], and compute the embedding similarity.

TF-IDF has been widely adopted in the field of information retrieval, and it reflects the importance of each word in a document. Some common words like $\text{tf}$ and $\text{torch}$ in the API signature are less informative, so their TF-IDF weights tend to be smaller. To address this issue, we set $\text{tf}.\text{maximum}(x, y) = \text{tf}.\text{minimum}(x, y)$. We map an API signature into its TF-IDF (term frequency-inverse document frequency) embedding [73], and compute the embedding similarity.
term frequency embedding \([e_1^j, e_2^j, \ldots, e_n^j] \), where \(e_j^i \) is the number of occurrences of word \(j \) in the API signature of \(S \). We further normalize it with the inverse document frequency for each word to get the TF-IDF embedding:

\[
Rep_{sig}(S) = \frac{[c_1^S, c_2^S, \ldots, c_n^S]}{\sum_{S \in A} \sum_{i=1}^{c_i}}
\]

The cosine similarity of two vectors is the cosine of the angle between them, and thus always belongs to the interval \([-1, 1]\]. The cosine similarity of two arbitrary vectors \(x, y\) is defined as follows:

\[
\cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|}
\]

We then compute the cosine similarity between the TF-IDF embeddings to be the signature similarity of two APIs \((S, T)\):

\[
Sim_{sig}(S, T) = \cos(Rep_{sig}(S), Rep_{sig}(T))
\]

Document Similarity. To complement the signature similarity, we further model the semantic similarity between API documents. First, we extract all API descriptions from the documentation, each of which is a one-sentence summary of an API given at the beginning of the document. For example, API `torch.vsplit` is described as "Splits input into multiple tensors vertically." The description succinctly and surgically states the expected input, the transformation applied, and the expected output. Next, we use Sentence-BERT [58] to encode these informative description sentences into semantically meaningful sentence embeddings. The Sentence-BERT encoder takes a natural language sentence as input and outputs a vector in high-dimensional space; it targets specifically at generating embeddings whose cosine similarity reflects the semantic textual similarity. We use \(Sdescription\) to denote the description for the API \(S\), and use \(SBEncoder\) to denote the Sentence-BERT encoder. For each API \(S\), we obtain its document embedding as:

\[
Rep_{doc}(S) = SBEncoder(S.description)
\]

For each API pair \((S, T)\), we compute the cosine similarity between their document embeddings as their document similarity:

\[
Sim_{doc}(S, T) = \cos(Rep_{doc}(S), Rep_{doc}(T))
\]

### 4.2 Invocation Synthesizer

In this phase, we leverage argument matching and template matching to synthesize the invocation code for each matched API pair. Note that the invocation code of source API is simply the code snippet that directly invokes the source API. Therefore, we will next focus on generating the target API invocation code.

**Argument Matching.** For each candidate API pair, DeepREL first synthesizes the invocation code based on API definitions. It maps the arguments of the source API to the arguments of the target API to synthesize the invocation code of the target API (with the arguments from the source API).

We transform the argument matching problem into a maximum weighted bipartite matching problem [71]. DeepREL generates the invocation code based on the best argument match. More formally, given the source API \(S\), the target API \(T\), and their argument lists \(S.args\) and \(T.args\), the corresponding bipartite graph is \(G = (L, R, E)\), where \(L = \{a \mid a \in S.args\}\), \(R = \{b \mid b \in T.args\}\) and \(E = \{(a, b) \mid a \in L, b \in R\}\). The weight of each edge \((a, b) \in E\) is the similarity \(Sim_{arg}(a, b)\) of \(a\) and \(b\), which is defined as:

\[
Sim_{arg}(a, b) = Sim_{name}(a, b) + Sim_{type}(a, b) + Sim_{pos}(a, b)
\]

The similarity \(Sim_{arg}(a, b)\) is determined by the names, potential types, and positions of the arguments. First, the similarity of the names is computed based on the following formula:

\[
Sim_{name}(a, b) = 1 - \frac{\text{Levenshtein}(a.name, b.name)}{\text{Max}(\text{Len}(a.name), \text{Len}(b.name))}
\]

where \(a.name\) is the name of argument \(a\). This similarity is based on the Levenshtein Distance [70] between the two names. Next, we compute the similarity of two type sets as:

\[
Sim_{type}(a, b) = \frac{|a.type \cap b.type|}{|a.type|}
\]

where \(a.type\) is the set of possible types of \(a\). If the set of types that argument \(b\) can take contains all possible types of \(a\), \(a\) is more likely to be mapped to \(b\) since all types of \(a\) are legal for \(b\).

We also compute the positional similarity of two arguments as:

\[
Sim_{pos}(a, b) = 1 - \frac{|a.idx - b.idx|}{\text{Max}(\text{Len}(S.args), \text{Len}(T.args))}
\]
the possible type set. The type set could be incomplete since it relies on the traces. When no traces cover a particular argument $b$ of a target API, the type set $b_{type}$ will be empty; thus, the type similarity between $b$ and any other argument $a$ will be $\text{Sim}_{type}(a, b) = 0$. Note that the argument matching algorithm still works in this case, with only the name and positional similarities being considered.

**Template Matching.** For some complex API pairs, DeepREL cannot leverage argument matching to generate the correct invocation code of the target API. Figure 7b shows the right invocation code between `scatter_nd` and `tensor_scatter_nd_add`. Obviously, argument matching fails to synthesize the complex invocation code of `tensor_scatter_nd_add`, in which case a matching template can be useful. A matching template is a code snippet elaborating a matched API pair. It presents an invocation of the target API, whose inputs are obtained from the arguments of the invocation of the source API. Figure 7a shows an example matching template (highlighted with underline) from the documentation of `tf.scatter_nd`. It suggests that invoking `scatter_nd` is equivalent to invoking `tensor_scatter_nd_add` with proper argument mappings as shown in Figure 7b.

To automatically detect such templates, DeepREL examines each code block in the documentation. Whenever a code snippet contains the invocation of another API, DeepREL extracts it as a potential candidate. Note that not every API pair has such templates. If DeepREL failed to extract any, the matching template will simply be None. For API pairs with a matching template, regardless of whether it is in top-K similar pairs, DeepREL synthesizes additional invocation code using the matching template as the target API invocation code.

### 4.3 API Match Verifier

In this phase, DeepREL runs the invocation code synthesized for each matched API pair over a set of verifying inputs to validate properties `Equivalence_{value}` and `Equivalence_{status}` (defined in Section 3). The verifying inputs are a collection of valid inputs for the source API. These inputs can be collected from the documentation, library tests, and existing DL models, so that they are representative of the valid input space of the source API. For example, in our implementation, we leverage state-of-the-art FreeFuzz [66], which can collect all such information fully automatically.

**Equivalence_{value}**. DeepREL first checks whether a candidate API pair satisfies the `Equivalence_{value}` property given the verifying inputs. If so, DeepREL accepts this matching pattern and marks it as `Equivalence_{value}`. Take the API pair in Figure 7b for example, `scatter_nd` always has the same output as `tensor_scatter_nd_add` for all verifying inputs. Hence, it is labeled as `Equivalence_{value}`.

**Equivalence_{status}**. If the API pair violates `Equivalence_{value}`, DeepREL further checks whether the source and target API have the same status given verifying inputs. If so, DeepREL accepts this API pair relation as `Equivalence_{status}`. For instance, the `Equivalence_{value}` property does not hold for the pattern shown in Figure 3 since the output of `AdaptiveAvgPool3d` is different from `AdaptiveMaxPool3d` over the input set. However, the `Equivalence_{status}` property holds as they always have the same status over the verifying inputs.

If an API pair is verified as `Equivalence_{value}` or `Equivalence_{status}`, it will be accepted by the API Match Verifier and further tested in the next fuzzing phase; otherwise, DeepREL rejects this pair.
Please note that it is crucial to have a set of representative verifying inputs. If the verifying inputs are not representative, the API Match Verifier could mistakenly accept a wrong API relation when the verifying inputs do not cover certain important regions of the possible input space. We will study such false positives in detail in our experimental study (Section 6.2). On the other hand, the API Match Verifier may also reject some true matched API pairs if the verifying inputs directly trigger real consistency bugs in this phase. Although it is hard to avoid such false negatives, our experimental results show that DeepREL detects 162 bugs for popular DL libraries fully automatically, demonstrating the effectiveness of this design.

4.4 API Fuzzer

In the last phase, DeepREL further leverages the verified API pairs to detect bugs for DL libraries. Specifically, DeepREL applies off-the-shelf mutation-based fuzzing techniques [66] to mutate the source API inputs for generating diverse inputs for differential testing. For Equivalence\textsubscript{value}, the source and target APIs are expected to have the same output. We can detect consistency bugs by comparing the results of the source and target APIs. For Equivalence\textsubscript{status}, the source and target APIs are expected to have the same status. Thus, we compare their statuses to detect bugs.

4.5 The Iterative Process

In order to cover more APIs by matched API pairs, DeepREL performs the above four phases iteratively until the fixed point or a given number of iterations $I$ (by default 10). In the API Match Verifier phase, if the target API is not covered in the current iteration and its invocation generated by the synthesizer has success status, we will add this invocation into the database (even when the corresponding API relation got falsified) and label the target API as “newly covered API”. After this iteration, if there is any newly covered API, DeepREL will re-run the framework with these newly covered APIs as the source APIs; otherwise, the fixed point has been reached, and DeepREL will terminate. It is worth mentioning that the entire iterative DeepREL approach is fully automated. For the newly covered APIs, the verifying inputs are also automatically “borrowed” from the source APIs’ valid inputs.

Figure 4 presents one example of this iterative process. In the first iteration, the API Match Verifier takes API pair \texttt{(torch.vsplit, torch.tensor_split)} as input, and queries the invocation database \texttt{1} to get a record for \texttt{torch.vsplit}. Invoking \texttt{torch.tensor_split} \texttt{2} results in success. Assuming that \texttt{torch.tensor_split} is not covered in the database at the beginning of the iteration, this successful invocation is then inserted into the database \texttt{3}. In the next iteration, this invocation record will be retrieved \texttt{4} to verify the matched API pairs with \texttt{torch.tensor_split} as the source API.

5 EXPERIMENTAL SETUP

In the experiments, we address the following research questions:

- **RQ1:** How effective is DeepREL in terms of API coverage?
- **RQ2:** What is the false positive rate of DeepREL?
- **RQ3:** How do different configurations affect DeepREL?
- **RQ4:** Can DeepREL detect real-world bugs?

Our experiments are performed on PyTorch 1.10 [52] and TensorFlow 2.7 [13], the latest stable release versions for the two most popular DL libraries (with 54K/162K stars on GitHub), which are the most widely studied DL libraries in prior DL library testing work [30, 54, 65, 66]. We run all experiments on a machine with an 8-core 2.20GHz Intel CPU, 16GB RAM, and Ubuntu 20.04.

5.1 Implementation

**API Matcher.** To find high-quality matched API pairs, we first use the bs4 Python package [2] to parse the documentation of all 7973 APIs from TensorFlow and PyTorch. We collect both API signatures and descriptions from the documentation. Before computing the TF-IDF embedding, we use the Snowball stemmer [55] to convert tokens into word stems. As for the document embedding, we use the SentenceTransformer Python package [12] and adopt the pre-trained model all-MiniLM-L6-v2 as our \texttt{SBEncoder}.

**Invocation Synthesizer.** For argument matching, we use the munkres Python package [10] (implementing the Kuhn–Munkres algorithm) to solve the maximum weighted bipartite matching problem. For template matching, we automatically search and extract the code snippets for matching templates from the documentation.

**API Match Verifier.** We verify the relation of candidate API pairs with a set of representative valid inputs. We obtain the valid inputs traced from various input sources used by state-of-the-art FreeFuzz [66], which include the documentation, developer tests, and 202 DL models, and are representative to verify the function of APIs. We feed a maximum of 100 valid invocations of the source API from the FreeFuzz database into both the source API and target API as the verifying inputs and check if they have consistent behaviors.

**API Fuzzer.** We leverage the fuzzing strategies of FreeFuzz to mutate all the valid inputs traced for each source API (within the FreeFuzz database) and run all the generated inputs (1000 for each source API following the default setting of FreeFuzz [66]) on both the source API and target API to detect consistency bugs.

5.2 Metrics

**# of Covered API.** Following prior work in DL library testing [66], we report the number of covered APIs. An API is covered by DeepREL if it is successfully invoked by API Match Verifier either as a source API or a target API (i.e., invocations with the Success status). Since DL libraries contain a large number of APIs, API coverage is an important metric of test adequacy.

**False Positive Rate.** If an API pair satisfies that 1) at least one of its invocation code is accepted by the API Match Verifier, and 2) at least one of its accepted invocation code is against its labeled property during the fuzzing phase, it is named an inconsistent API pair. The false positive rate for inconsistent API pairs is the proportion of detected inconsistent API pairs which are false alarms. It is commonly used in prior work on fuzzing or testing [22, 61, 80].

**# of Detected Bugs.** Bug finding is the ultimate goal for fuzzing, and thus we also report the number of distinct bugs DeepREL finds.

6 RESULT ANALYSIS

6.1 RQ1: Effectiveness in API Coverage

In this RQ, we aim to study the effectiveness of DeepREL in terms of covering more APIs with API relations. Table 1 shows the number of DL library APIs covered by DeepREL and state-of-the-art FreeFuzz. Column “#Total” presents the total number of APIs in DL
labeled tensors and is a low-level API compared to the high-level API to all TensorFlow ops. For example, \texttt{tf.pad} contains 1339 low-level APIs to provide direct access to the input into the output as they become available, in some situations parallel operations are equivalent APIs which pile a list of tensor operations and thus have the same functionality. Table 3 further presents a detailed distribution of the API pairs inferred by DeepREL based on whether the source/target APIs involve newly covered APIs. Column “Src” and Row “Tgt” present the categorization of the source and target APIs, respectively; Columns “Seed” and “New” indicate whether an API is from “seed APIs” covered by FreeFuzz or is newly covered by DeepREL. Out of all the 13098 API pairs verified by DeepREL, 1679 (902 + 777) only involve APIs covered by the original FreeFuzz, and all the remaining 11419 pairs involve newly covered APIs, which shows the importance of leveraging API relations to cover more APIs.

We also conduct a manual study to investigate why there are so many value-equivalent API pairs. Note that we do not look into status-equivalent API pairs because they are more intuitive (e.g., many APIs may share similar input parameter types and/or output behaviors). Since the number of verified value-equivalent API Pairs is huge, we select the set of equivalent API pairs explicitly specified in the documentation for our study.

Table 2 shows the number of verified API pairs detected by DeepREL. Columns “Equivalence\_value” and “Equivalence\_status” present the number of value-equivalent and status-equivalent API pairs accepted by the API Match Verifier respectively. DeepREL accepts 13098 API pairs in total, showing that such API relations are common in DL libraries. On PyTorch, DeepREL accepts more status-equivalent API pairs (2933) than value-equivalent (1357). The reason is that the latter relation is stricter than the former, and status-equivalent API pairs are more common. For example, in term of splitting a tensor, PyTorch provides a set of APIs: torch.split, torch.tensor_split, torch.vsplit (splits the tensor vertically), and torch.dsplit (splits the tensor depthwise). It is worth noting that TensorFlow has much more APIs than PyTorch, and DeepREL detects more value-equivalent API pairs than status-equivalent ones on TensorFlow. This is because TensorFlow contains lots of APIs for compatibility and low-level access operations and thus has higher functional overlap: (1) The tf.compat module \cite{tf} contains 2579 redundant APIs to support forward and backward compatibility across versions. For example, \texttt{tf.compat.v1.layers.conv2d} is an alias for \texttt{tf.layers.conv2d}, and it allows user to use the \texttt{conv2d} layer with TensorFlow v1 behavior in TensorFlow v2; (2) The \texttt{tf.raw_ops} module \cite{tf2} contains 1339 low-level APIs to provide direct access to all TensorFlow ops. For example, \texttt{tf.raw_ops.Pad} adds padding to tensors and is a low-level API compared to the high-level API \texttt{tf.pad} with the same functionality.

Table 3 further presents a detailed distribution of the API pairs inferred by DeepREL based on whether the source/target APIs involve newly covered APIs. Column “Src” and Row “Tgt” present the number of value-equivalent and status-equivalent API pairs because they are more intuitive (e.g., many APIs may share similar input parameter types and/or output behaviors). Since the number of verified value-equivalent API Pairs is huge, we select the set of equivalent API pairs explicitly specified in the documentation for our study.

In the fuzzing stage, DeepREL covers 2973 APIs, which is a huge improvement (157%) over FreeFuzz that covers only 1158 APIs. For example, there are totally 1592 PyTorch APIs, and DeepREL can cover 1071 APIs, 128% more than FreeFuzz. A large number of APIs are not covered by FreeFuzz because they are less frequently used and not covered by any of the three sources of FreeFuzz. Leveraging API relations, DeepREL can successfully invoke these APIs with their relational APIs’ inputs. The huge API coverage improvement demonstrates the potential of DeepREL.
Table 4: Classification of reasons for Equivalence\textsubscript{value} API pairs

<table>
<thead>
<tr>
<th>Reason</th>
<th># TP</th>
<th>% TP</th>
<th># PT</th>
<th>% PT</th>
<th>Example (S, T)</th>
<th>Diff. between S and T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of programming</td>
<td>460</td>
<td>95.44%</td>
<td>349</td>
<td>91.36%</td>
<td>(torch.det, torch.linalg.det)</td>
<td>S is an alias for T</td>
</tr>
<tr>
<td>Performance</td>
<td>10</td>
<td>2.07%</td>
<td>9</td>
<td>2.36%</td>
<td>(tf.parallel_stack, tf.stack)</td>
<td>S uses parallelism for efficiency.</td>
</tr>
<tr>
<td>Special cases</td>
<td>10</td>
<td>2.07%</td>
<td>10</td>
<td>2.07%</td>
<td>(tf.boolean_mask, tf.ragged.boolean_mask)</td>
<td>T extends S to ragged tensors.</td>
</tr>
<tr>
<td>Numerical stability</td>
<td>1</td>
<td>0.21%</td>
<td>1</td>
<td>0.15%</td>
<td>(torch.linalg.inv, torch.linalg.solve)</td>
<td>T is faster and more numerically stable.</td>
</tr>
<tr>
<td>Deprecation</td>
<td>1</td>
<td>0.21%</td>
<td>1</td>
<td>0.21%</td>
<td>(torch.qr, torch.linalg.qr)</td>
<td>S is deprecated.</td>
</tr>
</tbody>
</table>

Table 5: False positive rate of DeepREL

<table>
<thead>
<tr>
<th>Oracle</th>
<th># All</th>
<th>TP</th>
<th>FP</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyTorch Equivalence\textsubscript{value}</td>
<td>412</td>
<td>364</td>
<td>48</td>
<td>11.65%</td>
</tr>
<tr>
<td>Overall</td>
<td>1265</td>
<td>918</td>
<td>347</td>
<td>27.43%</td>
</tr>
<tr>
<td>TensorFlow Equivalence\textsubscript{value}</td>
<td>97</td>
<td>68</td>
<td>29</td>
<td>29.90%</td>
</tr>
<tr>
<td>Overall</td>
<td>480</td>
<td>277</td>
<td>183</td>
<td>39.78%</td>
</tr>
<tr>
<td>Total Equivalence\textsubscript{value}</td>
<td>509</td>
<td>432</td>
<td>77</td>
<td>15.13%</td>
</tr>
<tr>
<td>Total Equivalence\textsubscript{status}</td>
<td>1216</td>
<td>763</td>
<td>453</td>
<td>37.25%</td>
</tr>
<tr>
<td>Overall</td>
<td>1725</td>
<td>1195</td>
<td>530</td>
<td>30.72%</td>
</tr>
</tbody>
</table>

API relation when the set of inputs does not cover certain part of the possible input space and thus is insufficient to distinguish the different behaviors of the API pair. We can also observe that the FPR of Equivalence\textsubscript{value} is lower than Equivalence\textsubscript{status}. This is because that Equivalence\textsubscript{value} is stricter than Equivalence\textsubscript{status}, making it easier for API Match Verifier to reject wrong API relations over verifying inputs.

6.3 RQ3: Impacts of Configuration

In this RQ we analyze how different configurations affect the performance of DeepREL, including the API coverage, FPR, and running time. We focus on two hyper-parameters, $K$ (the number of matched pairs for each source API, discussed in Section 4.1) and $I$ (the number of iterations, discussed in Section 4.5). The default values for $K$ and $I$ are both 10 in DeepREL. To figure out the impact of $K$, we run our experiments with different $K$ values of 5, 10, 15, 20. We also run DeepREL for up to 10 iterations and show the impact of different $I$ values from 1 to 10. Figures 8 and 9 show the results, where the x axis is the iterations, and the y axis is the number of covered APIs, FPR, and the running time respectively. The results for different $K$ values are shown in different lines. We can observe that DeepREL terminates (i.e., reaching fixed points) in at most 8 iterations on PyTorch, and on TensorFlow it either terminates in at most 8 iterations, or only covers 2 new APIs in the last iteration.

The impact of $K$ is similar for PyTorch and TensorFlow. First, the number of covered APIs of $K = 10, 15, 20$ are close, all significantly higher than $K = 5$. Second, the FPR increases as $K$ increases. The reason is that less similar API pairs can incur more false positives. Third, the running time is increasing approximately linearly with $K$. Therefore, it is reasonable to set $K = 10$ in our default setting, as setting $K$ to be higher than 10 would bring marginal benefit in terms of API coverage, but degrade FPR and time cost.

For the impact of $I$, we have the following observations. First, as expected, the number of covered APIs increases at a lower and lower pace with the increase of $I$, and converges within 10 iterations. Second, the FPR generally increases with iteration. The reason is that source APIs in later iterations are typically target APIs from earlier iterations, and may have fewer and fewer valid inputs (since they may fail on some inputs from the original source APIs) for verifying inferred API pairs. The only exception is that FPR drops when $I$ increases to 2 with $K = 5$ or 10 on PyTorch. We look into it and observe that DeepREL can accidentally detect a large number of true positives in the 2\textsuperscript{nd} iteration (e.g., torch.Tensor. + APIs and their value-equivalent APIs torch.*). Third, the running time grows more rapidly at early iterations and more slowly later, which is consistent with the growth of the number of covered APIs.

We can also observe that the total running time of the default DeepREL is 12.8h for PyTorch and 26.3h for TensorFlow. Such cost is actually quite common for fuzzing techniques, e.g., various fuzzing techniques have been applied for 24h or even more [17, 40, 59, 67, 80] (including the recent LEMON work [65]).

6.4 RQ4: Bugs Detected

Table 6 presents the summary of real-world bugs detected by DeepREL for the studied DL libraries. Column “Total” shows the total number of bugs detected by DeepREL, and Columns “Value” and “Status” show the number of bugs detected with the Equivalence\textsubscript{value} and Equivalence\textsubscript{status} oracle respectively. We also present the number of bugs rejected by developers (Column “Rejected”), confirmed as previously unknown (Column “Confirmed”), and the number of previously unknown bugs that have already been fixed (Column “Fixed”). We can observe that DeepREL is able to detect 162 bugs in total, with only 7 rejected by developers, 106 confirmed by developers as previously unknown bugs (48 already fixed), and all others pending. Among those 106 confirmed bugs, only 9 can be found by Fuzzing Deep-Learning Libraries via Automated Relational API Inference ESEC/FSE ’22, November 14–18, 2022, Singapore, Singapore FreeFuzz, and none of them can be detected by CRADLE, AUDEE, or LEMON. Furthermore, we have also found 10 documentation bugs for PyTorch and 4 for TensorFlow during the experiment (these document bugs are not included in Table 6).

The bugs detected by DeepREL can also be categorized into single API bugs\footnote{The single API bugs are unexpected crashes caused by single buggy APIs. In Table 6, we categorize such bugs as “Value”/“Status” if they were detected when testing API pairs with the Equivalence\textsubscript{value}/Equivalence\textsubscript{status} oracle (although they can be detected without such relational API oracles).} and consistency bugs between relational API pairs. Table 7 shows the breakdown of the 106 confirmed bugs based on the two cases. For the 19 single API bugs detected by DeepREL, 9 are in the “seed APIs” (Column “Seed”), while 10 are bugs in the newly covered APIs (Column “New”), emphasizing the importance of leveraging API relations to cover more APIs. Meanwhile, all the remaining 87 ones are consistency bugs, demonstrating the effectiveness of using API relations as the oracle for DL library testing. More specifically, out of the 87 consistency bugs, 35 are inconsistencies between “seed APIs” (Column “Seed-only”) while 52 involve newly covered APIs (Column “Others”), further indicating that covering new APIs contributes to consistency bug detection.

Notably, DeepREL is able to detect 23 high-priority bugs for PyTorch (note that TensorFlow is not discussed here as it does not have such labels). These bugs are marked as “high priority” because...
they are critical and require urgent resolution. The reported "high priority" bugs are highly valued by PyTorch developers and have raised heated discussions. During the three months from our first bug issue (November 23, 2021) to March 1, 2022, there have been a total of 170 high-priority issues in the entire PyTorch issue-tracking system [11], of which we contributed 23 (13.5%), emphasizing the effectiveness of DeepREL.

We next present some example bugs detected by DeepREL:

### Out-of-Bounds Read (Equivalence value, ✓)

The bug shown in Figure 10 is detected in API pair torch.kthvalue(tensor, k) and tensor.kthvalue(k) via Equivalence value. The latter API is a method of torch.Tensor [5], which is the fundamental class in PyTorch. Given the same tensor (input) and argument (k), the returned results are not equal (Line 4). After debugging, we find that the returned results can actually be different across runs, indicating that it reads values from memory locations outside of user-controlled data! This bug is a silent error, and has a severe security implication: without proper range checking of k, users may be able to read data outside of the allocated memory bounds (i.e., out-of-bound read). This bug is labeled as "high priority" and fixed immediately.

```python
input = torch.tensor([0,1,2,3,4])
result1 = torch.kthvalue(input,k=6)
result2 = input.kthvalue(k=6)
torch.testing.assert_close(result1, result2) # fail
```

### Inconsistent Check (Equivalence status, ✓)

Figure 11 shows a bug for AdaptiveAvgPool3d and AdaptiveMaxPool3d in PyTorch (found via Equivalence status). With exactly the same input tensor (tensor) and argument (output_size), AdaptiveAvgPool3d runs without exception (Line 3) but AdaptiveMaxPool3d throws RuntimeError: Trying to create tensor with negative dimension (Line 4). After inspection, we find that AdaptiveAvgPool3d lacks checking for dimensions with negative integers. This can be disastrous since users may inadvertently introduce a bug into their model, but no warning/exception is raised. This bug has also been fixed.

### Wrong Computation (Equivalence value, ×)

The bug shown in Figure 12 is detected by Equivalence value oracle for API pair torch.std_mean and torch.mean. Both of them could compute the mean of all elements in the input tensor. However, DeepREL has found that for certain input tensors, torch.std_mean has a different returned value from torch.mean. However, the developers said these two APIs are not expected to output the same mean and rejected
input = torch.tensor([[0.5786, 0.1719, 0.3760, 0.2939, 0.3984],
[0.5361, 0.7704, 0.0765, 0.0903, 0.0483]], dtype=float16)
result2 = torch.mean(input) # 0.4082
result1 = torch.std_mean(input) # 0.4080

We have talked about related work on DL library testing in Section 2. Widely used DL libraries, TensorFlow and PyTorch. Last but not least, the threats to construct validity mainly lie in the metrics used. For example, SBES [26, 43] synthesizes sequences of method invocations that are equivalent to a target method based on the observed dynamic program behaviors. The recent state-of-the-art technique MeMo [16] mainly targets mature open-source Java projects and automatically derives metamorphic relations from natural language specifications. Meanwhile, MeMo can only infer metamorphic relations that are well documented and lacks an invocation synthesizer or API relation verifier; thus the effectiveness of MeMo strictly depends on documentation completeness, correctness, and preciseness. In fact, very few metamorphic relations in DL libraries are explicitly described with informative sentences (with code templates) and can be directly translated into valid oracles. We substituted the first three phases of DeepREL with MeMo, and found that it can infer only 17/28 API relations for TensorFlow/PyTorch (compared to 8808/4290 relations found by DeepREL), and can at most detect 8 (including 6 found by MeMo and 2 found by FreeFuzz on the APIs involved in the MeMo inferred relations) of the 106 confirmed bugs detected by DeepREL.

ACKNOWLEDGMENTS
We appreciate the insightful comments from the anonymous reviewers. This work was partially supported by National Science Foundation under Grant Nos. CCF-2131943 and CCF-2141474, as well as the Google Research Scholar Program.