

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and
Information Systems

School of Computing and Information Systems

11-2022

AutoPruner: transformer-based call graph pruning

Cong Thanh LE

Singapore Management University, tlecong@smu.edu.sg

Hong Jin KANG

Singapore Management University, hjkang.2018@phdcs.smu.edu.sg

Truong Giang NGUYEN

Singapore Management University, gtnguyen@smu.edu.sg

Stefanus AGUS HARYONO

Singapore Management University, stefanusah@smu.edu.sg

David LO

Singapore Management University, davidlo@smu.edu.sg

See next page for additional authors

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research



Part of the [Software Engineering Commons](#)

Citation

LE, Cong Thanh; KANG, Hong Jin; NGUYEN, Truong Giang; AGUS HARYONO, Stefanus; LO, David; LE, Xuan-Bach D.; and THANG, Huynh Quyet. AutoPruner: transformer-based call graph pruning. (2022).

Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Singapore, 2022 November 14-18. 520-532.

Available at: https://ink.library.smu.edu.sg/sis_research/7740

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylids@smu.edu.sg.

Author

Cong Thanh LE, Hong Jin KANG, Truong Giang NGUYEN, Stefanus AGUS HARYONO, David LO, Xuan-Bach D. LE, and Huynh Quyet THANG

AUTOPRUNER: Transformer-Based Call Graph Pruning

Thanh Le-Cong
Hong Jin Kang
Truong Giang Nguyen
Stefanus Agus Haryono

David Lo
Singapore Management University
Singapore, Singapore

Xuan-Bach D. Le
University of Melbourne
Melbourne, Victoria, Australia

Quyet Thang Huynh
Hanoi University of Science and
Technology
Hanoi, Vietnam

ABSTRACT

Constructing a static call graph requires trade-offs between soundness and precision. Program analysis techniques for constructing call graphs are unfortunately usually imprecise. To address this problem, researchers have recently proposed *call graph pruning* empowered by machine learning to post-process call graphs constructed by static analysis. A machine learning model is built to capture information from the call graph by extracting structural features for use in a random forest classifier. It then removes edges that are predicted to be false positives. Despite the improvements shown by machine learning models, they are still limited as they do not consider the source code semantics and thus often are not able to effectively distinguish true and false positives.

In this paper, we present a novel call graph pruning technique, AUTOPRUNER, for eliminating false positives in call graphs via both statistical semantic and structural analysis. Given a call graph constructed by traditional static analysis tools, AUTOPRUNER takes a Transformer-based approach to capture the semantic relationships between the caller and callee functions associated with each edge in the call graph. To do so, AUTOPRUNER fine-tunes a model of code that was pre-trained on a large corpus to represent source code based on descriptions of its semantics. Next, the model is used to extract semantic features from the functions related to each edge in the call graph. AUTOPRUNER uses these semantic features together with the structural features extracted from the call graph to classify each edge via a feed-forward neural network. Our empirical evaluation on a benchmark dataset of real-world programs shows that AUTOPRUNER outperforms the state-of-the-art baselines, improving on F-measure by up to 13% in identifying false-positive edges in a static call graph. Moreover, AUTOPRUNER achieves improvements on two client analyses, including halving the false alarm rate on null pointer analysis and over 10% improvements on monomorphic call-site detection. Additionally, our ablation study and qualitative analysis show that the semantic features extracted by AUTOPRUNER capture a remarkable amount of information for distinguishing between true and false positives.

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 the author(s) 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 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9413-0/22/11...\$15.00

<https://doi.org/10.1145/3540250.3549175>

CCS CONCEPTS

- **Software and its engineering** → **Automated static analysis**;
- **Computing methodologies** → **Machine learning**.

KEYWORDS

Call Graph Pruning, Static Analysis, Pretrained Language Model, Transformer

ACM Reference Format:

Thanh Le-Cong, Hong Jin Kang, Truong Giang Nguyen, Stefanus Agus Haryono, David Lo, Xuan-Bach D. Le, and Quyet Thang Huynh. 2022. AUTOPRUNER: Transformer-Based Call Graph Pruning. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '22)*, November 14–18, 2022, Singapore, Singapore. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3540250.3549175>

1 INTRODUCTION

Call graphs construction is crucial for program analyses. Call graphs capture invocations between functions of programs [7, 43]. An ideal call graph is (1) sound, i.e., it does not miss any true invocations, and (2) precise, i.e., it does not produce any false positives. However, even for small programs, constructing a sound and precise call graph is difficult [1]. A call graph analysis should make reasonable trade-offs between soundness and precision. Unfortunately, recent work [51] has found that widely used tools such as WALA [12] or Petablox [30] construct imprecise call graphs; up to 76% of edges in call graphs constructed by WALA are false positives.

To address these issues, researchers [4, 30, 48] have proposed to improve pointer analysis, which is the core of many call graph constructions algorithms, by improving context-sensitivity or flow-sensitivity of the analysis. Unfortunately, a perfect pointer analysis is generally not possible [42]. Specifically, pointer analyses usually face an expensive trade-off between scalability and precision [27]. For example, a context-sensitive analysis by WALA only reduces 8.6% false positives rate over a context-insensitive analysis while incurring a large performance overhead [51].

A recent approach, which we refer to as CGPRUNER [51], achieved a breakthrough in improving the quality of call graphs. Instead of directly improving pointer analysis, CGPRUNER performs call graph *pruning* as a post-processing technique on a call graph constructed through static analysis. Using machine learning techniques, the call graph pruner removes false positives in a call graph. Specifically, CGPRUNER first extracts a set of *structural* features from the call graph, e.g., the number of outgoing edges from the call-site and the in-degree of the callee. It then leverages a learning model, i.e.,

Random Forest, to predict if a caller-callee edge is a false positive (i.e., the caller does not invoke the callee in reality). The call graph is updated by removing edges that are predicted to be false positives. Since the cost of generating predictions is low, a machine learning-based approach does not incur a significant performance overhead. Their experiments show that CGPRUNER successfully improves over traditional call graph analysis by producing call graphs that eliminate a large number of false positives. However, CGPRUNER is still limited as it does not consider source code semantics and thus is not able to distinguish true and false positives effectively.

In this paper, we propose a novel technique, AUTOPRUNER, that incorporates both structural and statistical semantic information to prune false positives in call graphs effectively. AUTOPRUNER combines structural features extracted from a call graph with semantic features extracted from the source code of the caller and callee functions. Similar to CGPRUNER, AUTOPRUNER uses handcrafted structural features. However, different from CGPRUNER, AUTOPRUNER automatically extracts semantic features via deep learning. AUTOPRUNER thus faces a unique challenge on how to use deep learning to automatically learn from a limited dataset. To address this, we leverage recently proposed transformer models of code, i.e., CodeBERT [11], that has been pre-trained on a corpus containing millions of code functions. As our task differs from CodeBERT’s pre-training task, AUTOPRUNER first fine-tunes CodeBERT such that it captures the statistical relationships between caller and callee functions, learning to distinguish between true and false positive edges based on their source code. Next, AUTOPRUNER leverages the fine-tuned model to extract *semantic features* of each edge, based on the source code of the caller and callee functions. Each edge then has an embedding that represents the semantic relationship between the caller and the callee. For each embedding, AUTOPRUNER combines it with the handcrafted structural features to obtain a representation for each edge. Based on this representation, AUTOPRUNER employs a neural classifier to classify each edge in a call graph as a true or false positive.

We evaluate AUTOPRUNER on call graphs produced by three well-known tools, i.e., WALA [12], Doop [4], and Petablox [30], for real-world programs taken from the NJR-1 benchmark [36] in the same setting of CGPRUNER [51]. We compare the call graph generated by AUTOPRUNER against multiple baselines, including the original call graphs produced by WALA, Doop and Petablox, the call graphs pruned by state-of-the-art approach CGPRUNER, as well as a graph neural network model, that applies deep learning to the call graphs. The latter two baselines consider only structural information. Our experiments show that AUTOPRUNER improves over the state-of-the-art approach by up to 13% in F-measure. Our experiments demonstrate that the use of the semantic features extracted by the transformer-based model enables AUTOPRUNER to outperform approaches that consider only structural information.

We investigate the effect of pruned call graphs produced by AUTOPRUNER on client analyses, which take the call graphs as input to perform other analyses on the programs. We investigate two client analyses: null pointer analysis and monomorphic call-site detection. On null pointer analysis, call graphs pruned by our approach AUTOPRUNER lead to significantly reduced false alarm rate of only 12% while the call graphs pruned by CGPRUNER [51] and the call graphs constructed by WALA [12] have false alarm

rates in null pointer analysis of 23% and 73%. On monomorphic call-site detection, AUTOPRUNER improves over CGPRUNER by over 8% in terms of F-measure.

To better understand AUTOPRUNER, we also perform qualitative analysis on its performance. We leverage t-SNE [52] to visualize the embedding of the call graph edges in a two-dimensional space. We find that the semantic features can separate the true and false-positive edges, demonstrating that AUTOPRUNER captures a remarkable amount of information from the source code associated with each call graph edge.

In summary, we make the following contributions:

- We introduce AUTOPRUNER, a novel call graph pruner that uses both code and structural feature to identify false-positive edges in a call graph.
- We empirically demonstrate that pruned call graph produced by our approach can help analysis tool significantly improve the false alarm rate and F-measure. Notably, in the analysis client of null pointer analysis, AUTOPRUNER leads to over 150 more reported warnings while decreasing false alarm rate from 73% to 12%.
- We perform an ablation study and qualitative analysis to better understand our approach. Our analysis validates the use of our proposed approach for call graph pruning.

The paper is structured as follows: Section 2 introduces the background of our work. Section 3 describes our proposed approach. Section 4 presents our experimental setup and results. Section 5 discusses our qualitative analysis and threats to validity. Section 6 covers related work. Finally, Section 7 concludes and describes future directions.

2 BACKGROUND

In this section, we discuss a motivating example. Next, we present the formal formulation of the call graph pruning problem and introduce Transformer-based models of code and CodeBERT [11].

2.1 Motivating Example

In Figure 1, we present a motivating example to motivate our approach and demonstrate the limitations of CGPRUNER that uses only structural features to prune call graphs. The source code of the `acceptState` function contains a call to the `parse` function. In the original unpruned call graph, the `acceptState` node is connected to multiple `parse` nodes of classes that implement the `MathExpressionParser` interface, e.g. the interface is implemented by `EndOfExpressionParser`, `FunctionLeftParenthesisParser`, and other classes that also override `accept`. At runtime, the `acceptState` function is invoked multiple times with different values of `state`, resulting in calls between `acceptState` and the multiple `parse` nodes. These edges in the call graph are true edges as they represent calls that occur at runtime.

Due to a large number of outgoing edges from the same call site, the local structure of each `acceptState` to `parse` edge in the call graph resembles edges that are false positives. A structural-only approach such as CGPRUNER, therefore, incorrectly prunes all the edges between the `acceptState` node and the `accept` node in the call graph. This highlights the limitation of considering only the structural features of the call graph and motivates the need

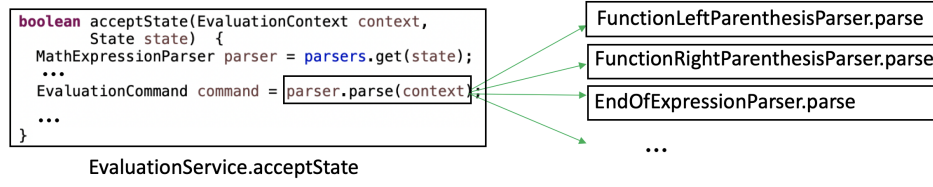


Figure 1: The `parser.parse` call depends on the parameter `state`. `parsers.get` returns an object of one of multiple classes implementing `MathExpressionParser`. Throughout the course of the program execution, `acceptState` can be invoked with all possible values of `state`. Therefore, despite the large number of outgoing edges from the same call site, all edges to the `parse` nodes from `acceptState` in the statically computed call graph are true edges. However, the large number of outgoing edges is a feature used to prune false positives, which leads to the incorrect pruning of the edges related to the `parse` calls.

for guiding call graph pruning with the semantics of the source code. From analyzing the source code, we can correctly identify that a large number of outgoing edges are possible, as the specific `parse` call depends on the parameters of the function. Indeed, AUTOPRUNER correctly leaves the edges unpruned due to its use of semantic features extracted from the source code by CodeBERT.

2.2 Call Graph Pruning

2.2.1 Problem Formulation. In this work, we formulate the call graph pruning problem as below.

Input: A static call graph $\mathcal{G} = (V, E)$ is a directed graph constructed by a static analysis tool, where V is the set of program’s functions identified by a function signature and E is the set of edges, i.e., function calls, in the call graph. Each edge in E is a tuple of (caller, callee, offset), where caller is the calling function, callee is the called function, and offset is the call-site in caller.

Output: A pruned call graph $\mathcal{G}' = (V', E')$, where $V' = V$ and $E' \subseteq E$.

To address this problem, we aim to train a binary classifier C , which can classify each edge in a call graph \mathcal{G} as **true positive**, i.e., the edge represents a true call, or **false positive**, i.e., the edge does not represent a true call. Using the classifier’s output, we prune the call graph following Algorithm 1. We use the classifier C to classify each edge in a call graph. Edges classified as false positives are pruned, while edges classified as true positives are retained.

Algorithm 1: Call-graph Pruner

Input: Call Graph $\mathcal{G} = (V, E)$, Classifier C

Output: Pruned Call Graph $\mathcal{G}' = (V', E')$

```

1  $\mathcal{G}' \leftarrow \mathcal{G}$ 
2 foreach  $e$  in  $\mathcal{G}$  do
3    $p \leftarrow C(e)$  ▷ prediction of binary classifier
4   if  $p == \text{False-positive}$  then
5      $E' = E' \setminus \{e\}$  ▷ remove edge from call graph
6   end
7 end
8 return  $\mathcal{G}'$ 

```

2.2.2 State-of-the-art. To prune call graphs, Utture et. al. [51] recently proposed CGPRUNER, which uses a machine learning model,

a random forest classifier, based on 11 structural features extracted from a call graph for pruning edges. Their experiments demonstrate that CGPRUNER could successfully boost the precision of call graphs from 24% to 66% and reduce the false positive rate in client tool from 73% to 23%. The approach, however, also substantially reduces the call graph’s recall.

CGPRUNER relies only on structural features, which can not distinguish false-positive and true-positive edges that share the same characteristics of structure (as mentioned in Section 2.1). Similar to CGPRUNER, our approach, i.e. AUTOPRUNER, employs a machine learning model to prune call graphs. However, unlike CGPRUNER, we use the semantic information from the source code of both the caller and callee function of an edge in the call graph. The information from the source code enables our approach to distinguish true-positive edges from false positive edges (later demonstrated in Section 5.1).

2.3 Transformer Models of Code and CodeBERT

Transformer models [53] are deep learning models based on an encoder-decoder architecture. Transformer models employ the attention mechanism and have achieved remarkable performance in field of Natural Language Processing (NLP) [5, 22, 39, 53, 56, 60]. NLP models have also been employed for source code-related tasks, as source code has been found to exhibit characteristics, such as repetitiveness and regularity, similar to natural language [17].

Recently, CodeBERT [11] was proposed as a Transformer model for source code. A CodeBERT model is pre-trained on a large corpus, containing over 6 million functions and 2 million pairs of comment-function. As input, CodeBERT can be given a pair of data (e.g., source code and a code comment that describes the semantics of the source code) to learn statistical relationships between the pair of data. CodeBERT is pre-trained by two tasks, Masked Language Modeling (MLM) and Replaced Token Detection (RTD). In the MLM task, given an input sequence with a single token, CodeBERT has to predict the value of the masked token. In the RTD task, given an input sequence where some tokens are replaced with alternative tokens, CodeBERT detects the replaced tokens. Previous studies have demonstrated the effectiveness of CodeBERT in multiple tasks, including the capability for CodeBERT to be fine-tuned for tasks that it was not initially trained for [11, 57]. Prior studies have been built on top of CodeBERT for automating various tasks that

require an understanding of program semantics, e.g., type inference [16, 21, 37], program repair [31], etc. Motivated by these success cases, our proposed solution (AutoPruner) built upon CodeBERT for another task, namely call-graph pruning. Our solution tries to capture the semantic relevance of code in the caller function to that of the callee function to differentiate between true and false edges in a call graph. Different from CGPRUNER, AutoPruner analyzes the source code of the functions while CGPRUNER ignores them. Moreover, AutoPruner considers both caller and callee functions while other pure program analysis methods analyze only the caller function.

3 METHODOLOGY

Figure 2 illustrates the overall framework of AUTOPRUNER. Before AUTOPRUNER can be applied, it has to be fine-tuned and trained. First, in the Fine-tuning phase, we fine-tune a pre-trained CodeBERT model of code to enable it to extract the *semantic features* of the edges in a call graph. Next, in the Training phase, we use the fine-tuned model to extract *semantic features* for the edges in the call graph. The semantic features are then combined with *structural features* extracted from the call graph to construct the representation of each edge. Then, we train a feed-forward neural network classifier to identify false positive edges. Afterward, in the Application phase, AUTOPRUNER can be used as a call graph pruner for post-processing call graphs to be used in other program analyses.

3.1 Fine-tuning

In this phase, we fine-tune CodeBERT. Before CodeBERT can be used for a task different from its pretraining task, it has to be fine-tuned to adapt its weights for the new task. CodeBERT takes a pair of data as input, and in AUTOPRUNER, we pass the source code of the caller and callee functions associated with each edge as input. Specifically, AUTOPRUNER uses a preprocessor that extracts the source code of the caller and callee functions, constructing a sequence of input tokens that matches the input format expected by CodeBERT. Then, the sequences are input into CodeBERT for fine-tuning. Below, we explain each component of the fine-tuning phase in detail.

3.1.1 Pre-processing. The pre-processing step (① in Figure 2) produces the input sequences to CodeBERT, enabling it to learn a representation of the semantic relationship between the source code of the caller and callee functions. Initially, from the call graph constructed by a static analysis tool (e.g., WALA or Doop), an edge in the call graph is characterized by a pair of function signatures identifying the caller and callee functions associated with it. We use *java-parser*¹ to extract the source code of both the caller and callee functions. Particularly, we parse the source code to obtain the method descriptors for every methods and matches them against the output of existing CG generators, which identifies methods using their descriptors. This allows us to link the source code to the methods in the call graph. Then, we use CodeBERT tokenizer² to tokenize the source code. Finally, following the input format of

CodeBERT, we construct an input sequence that encodes the source code of the caller and callee functions in the form of:

$$[CLS]\langle \text{caller's source code} \rangle [SEP]\langle \text{callee's source code} \rangle [EOS] \quad (1)$$

where [CLS], [SEP], and [EOS] are tokens separating the pair of data, as required by CodeBERT.

3.1.2 CodeBERT Fine-tuning. As the pre-training tasks of CodeBERT, (i.e., the Masked Language Modeling and Replaced Token Detection tasks) differs from our task (i.e., identifying false positive edges), we perform a fine-tuning step to adapt the pre-trained CodeBERT model for our task, following the common practice in transfer learning [50, 59] and other applications of CodeBERT [3, 31, 57, 58]. This step aims to transfer the knowledge based on the pre-training task associated with an extremely large amount of data onto our task where collecting data is expensive (as obtaining the ground truth labels in our task requires careful human analysis and the execution of test cases).

We fine-tune the CodeBERT model directly on the training dataset of our task of identifying false positive edges in the call graph. Specifically, we feed input sequences in the input format of CodeBERT, obtained from pre-processing step into the CodeBERT model. Next, the model extract features from the input sequences. Then, we pass the extracted features into a fully connected layer to classify each edge in the call graph into true and false positive edges (② in Figure 2).

During the fine-tuning phase, the parameters of the CodeBERT model are updated by Adam optimizer [23] to minimize the Cross-Entropy Loss. After this fine-tuning phase, we freeze all parameters of the CodeBERT model. In the subsequent phases, AUTOPRUNER uses the fine-tuned CodeBERT model to extract semantic features from the source code of the caller and callee functions associated with each edge.

3.2 Training

In the training phase, our objective is to train the binary classifier that predicts if a given edge is a true positive or false positive (the Classifier C in Algorithm 1). To this end, we construct the representation of an edge in a call graph by extracting and combining features of both types: semantic features (extracted from the source code by fine-tuned language model) and structural features (extracted from the call graph). Then, we train a neural classifier to predict whether an edge is true or false positive. Below, we explain each component of the pipeline.

3.2.1 Feature Extraction. We extract into a feature set (③ in Figure 2) the two types of features as follows:

- **Semantic features.** The semantic features are extracted from the source code of caller and callee functions using our fine-tuned CodeBERT model. To capture this information, we first apply the same pre-processing step as described in the Fine-tuning step (④ in Figure 2). Next, the fine-tuned CodeBERT model extracts a high-dimensional vector that encodes the statistical relationship between the caller and

¹<https://javaparser.org/>

²<https://huggingface.co/microsoft/CodeBERT-base/tree/main>

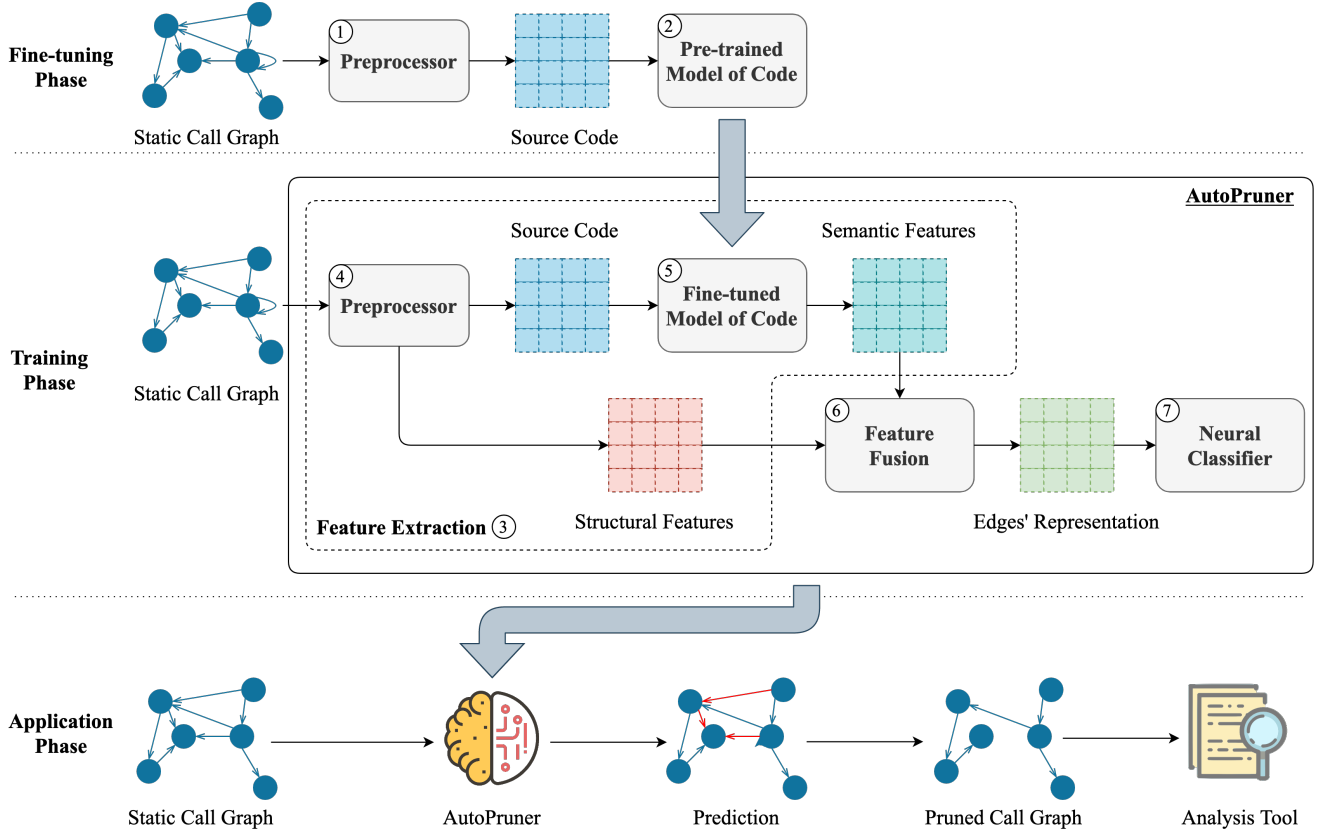


Figure 2: The overview of AUTOPRUNER

Table 1: Types of structural features

Feature	Description
<i>src-node-in-deg</i>	number of edges ending in caller
<i>src-node-out-deg</i>	number of edges out of caller
<i>dest-node-in-deg</i>	number of edges ending in callee
<i>dest-node-out-deg</i>	number of edges out of callee
<i>depth</i>	length of shortest path from main to caller
<i>repeated-edges</i>	number of edges from caller to callee
<i>L-fanout</i>	number of edges from the same call-site
<i>node-count</i>	number of nodes in call graph
<i>edge-count</i>	number of edges in call graph
<i>avg-degree</i>	average <i>src-node-out-deg</i> in call graph
<i>avg-L-fanout</i>	average <i>L-fanout</i> value in call graph

callee function ((5) in Figure 2). As a result, we obtain semantic features of an edge in call graph as follows:

$$f_{sem} = \langle v_1^{sem}, v_2^{sem}, \dots, v_{k_c}^{sem} \rangle \quad (2)$$

where $k_c = 768$ is the embedding dimension of CodeBERT.

- **Structural features.** The structural feature captures graphical information related to each edge. The features include metrics about the neighborhood of the edge (local information) or the entire call graph (global information). We use

the same features proposed by Utture et al. [51]. Detailed information of the features is presented in Table 1. We represent the structural features of an edge in a call graph as follows:

$$f_{struct} = \langle v_1^{struct}, v_2^{struct}, \dots, v_{k_s}^{struct} \rangle \quad (3)$$

where $k_s = 22$ is the number of structural features. Based on the work by Utture et al. [51], there are two features of each type listed in Table 1, one for transitive calls and one for direct calls, so we have 22 structural features.

3.2.2 Feature Fusion. In this step, we combine semantic features and structural features of each edge in the call graph into a final representation. We first use one fully connected layer for each feature ((6) in Figure 2). Then, the output of these layers are concatenated to produce the final representation. More formally,

$$f'_{sem} = FCN_{k_c \times h}(f_{sem}) \quad (4)$$

$$f'_{struct} = FCN_{k_s \times h}(f_{struct}) \quad (5)$$

$$f = \langle f'_{sem}, f'_{struct} \rangle \quad (6)$$

where $FCN_{m \times n}$ denotes a fully connected layer that takes a m -dimensional input and outputs a n -dimensional vector. We set h , which is the size of the hidden feature vector, as 32. k_c and k_s is the size of semantic and structural feature vector, respectively.

3.2.3 Neural Classifier. Given the representation of an edge obtained from feature extraction, we obtain a score that approximates the probability that an edge is a true positive.

The score is computed through a feed-forward neural network (⑦ in Figure 2) consisting of one hidden layer and one output layer. More formally,

$$f' = FCN_{2h \times 2}(f) \quad (7)$$

$$prob = OutLayer(f') \quad (8)$$

where, $FCN_{m \times n}$ denotes a fully connected layer inputs a m -dimensional vector and outputs a n -dimensional one, $OutLayer$ is a softmax function [13]. f and f' are the edge representation and output features, respectively. $prob = \{prob_{FP}, prob_{TP}\}$ is the output probabilities, where $prob_{FP}$ and $prob_{TP}$ is the probability that an edge is false positive and true positive, respectively. An edge with $prob_{TP}$ larger than $prob_{FP}$ is considered as a true positive. Otherwise, the edge are considered as a false positive.

During the training phase, the parameters of AUTOPRUNER except CodeBERT's parameters are updated by Adam optimizer [23] to minimize the cross-entropy loss [13].

3.3 Application

After the Training Phase, AUTOPRUNER can now be deployed for use as a call graph pruner. Given a call graph generated by a static analysis tool, AUTOPRUNER preprocesses the call graph and extracts semantic and structural features. Based on these features, the neural classifier produces predictions for each edge in the call graph. Using the predictions of the neural classifier, AUTOPRUNER removes the edges predicted to be false positives, and outputs an improved call graph with fewer false positives.

AUTOPRUNER can be integrated into other static analyses (i.e., client analyses) that takes a call graph as input. Since call graphs pruned by AUTOPRUNER have fewer false positives compared to the original ones, the performance of the client analyses should improve given the more precise call graphs.

4 EMPIRICAL EVALUATION

4.1 Research Question

Our evaluation aims to answer the following research questions:

RQ1: *Is AUTOPRUNER effective in pruning false positives from static call graphs?* This research question concerns the ability of AUTOPRUNER in identifying false positive edges in a static call graph. To evaluate our approach, we evaluate AUTOPRUNER on a dataset of 141 real-world programs in NJR-1 dataset [36] in terms of Precision, Recall, and F-measure. We compare our approach to multiple baselines, including the state-of-the-art technique, CGPRUNER [51], as well as a graph neural network, and the original call graphs produced by static analysis tools.

RQ2: *Can AUTOPRUNER boost the performance of client analyses?* This research question investigates the impact of pruned call graphs produced by AUTOPRUNER on client analysis. To answer this question, we use pruned call graph as input for client analyses and investigate its performance compared to original call graph and CGPRUNER [51] on two client analyses used by CGPRUNER: null-pointer analysis and monomorphic call-site detection.

RQ3: *Which components of AUTOPRUNER contributes to its performance?* AUTOPRUNER uses multiple types of features, including the semantic features extracted from both the caller and callee functions, and the structural features extracted from the call graph. We investigate the contribution of each feature in an ablation study by dropping each type of feature and observing the change in AUTOPRUNER's performance.

4.2 Experimental Setup

4.2.1 Dataset. To evaluate effectiveness of our approach, we use a dataset of 141 programs, initially constructed by Utture et al. [51], from the NJR-1 benchmark suite [36]. We follow the same experimental setting as prior work [51].

This dataset of programs was curated by Utture et al. [51] based on the criteria that each program has over 1,000 functions, has over 2,000 call graph edges, and has over 100 functions that are invoked at runtime, and has a high overall code coverage (68%). In total, the dataset comprises over 860,000 call graph edges. The ground-truth label of each edge was obtained based on instrumentation and dynamic analysis [51]. We use 100 programs as our training set and the remaining programs for the test set.

4.2.2 Evaluation Metrics. We estimate the quality of a static call graph using standard evaluation metrics: Precision, Recall, and F-measure, which are defined as follows:

$$Precision = \frac{|S \cap G|}{|S|} \quad (9)$$

$$Recall = \frac{|S \cap G|}{|G|} \quad (10)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

where S and G are the edge set in the static call graph and ground-truth respectively.

Among these evaluation metrics, Precision is the proportion of edges in call graph that are true calls. A high Precision is desirable for reducing developer effort in inspecting false positives [9]. Recall refers to the proportion of true edges that are retained in call graph. Finally, we consider F-measure, which is the harmonic mean of Precision and Recall. We use F-measure as a summary statistic to capture the tradeoff between Precision and Recall.

Following previous work [51], we compute the average Precision, Recall, and F-measure for the evaluation set by taking the mean over Precision, Recall, and F-measure of individual programs. We also report the standard deviation of each metric.

4.2.3 Client analyses. A better static call graph should lead to practical improvements on client analyses using the call graph. To assess the effect of the improvements to the call graph from AUTOPRUNER, we run experiments on two client analyses, null pointer analysis, and monomorphic call-site detection, using the call graph produced by WALA. To perform a direct comparison, the client analyses selected are the same as those considered by Utture et al. [51]. For each client analysis, we compare AUTOPRUNER against the baseline that produced the best call graph.

Null pointer analysis. In the first client analysis, we use analysis by Hubert et al. [18], which is implemented in Wala [12], to detect possible null pointer dereferences related to uninitialized

instance fields based on the input static call graph. This analysis is context-insensitive and field-insensitive. Improving the analysis would reduce the amount of developer effort spent inspecting false alarms, which is known to be a barrier to the adoption of bug detection tools based on static analysis [20]. For this client analysis, we refer to incorrect warnings reported as “false alarms” to distinguish them from “false positives” from the call graph construction. This analysis is independently checked by two human annotators to manually determine if the warnings are false alarms. In cases where the two annotators disagree on the decision, we involve a third annotator, an author of the paper, for a discussion to reach a consensus. We report the total number of reported warnings and the false alarm rate of the null pointer analyzer. As the ground-truth labels of all warnings are not known, we are not able to compute Recall and, therefore, we do not compare the approaches on F-measure.

Monomorphic call-site detection. In the second client analysis, the static call graph is used to detect monomorphic call sites. A call site is monomorphic when only one concrete function can be invoked. The detection of monomorphic call sites is useful, for example, for function inlining to reduce the runtime cost of function dispatch [51]. A better call graph would allow us to safely inline more functions, improving the performance of programs. The ground-truth of this analysis is determined by running the analysis on the ground-truth call graph. We report the precision, recall, and F-measure of the monomorphic call site detector.

Table 2: The classification threshold values of CGPRUNER for different static call graphs

Call Graph	WALA	Doop	Petablox
Balanced point	0.4500	0.4028	0.4279
Default	0.4500	0.5000	0.5000

4.2.4 Baselines. To assess the effectiveness of AUTOPRUNER, we compare our approach with the following baselines:

- the *original* call graphs are call graphs constructed by static analysis tools. In this work, we consider the 0-CFA static call graphs constructed by three standard static analysis tools, WALA [12], Doop [4], and Petablox [30]. The choice of these three analysis tools follows the previous work [51] for a fair comparison.
- a *random* baseline that randomly removes $N\%$ of edges in a call graph. For a fair comparison, we set N as the percentage of edges that are removed by AUTOPRUNER.
- CGPRUNER [51] is the state-of-the-art technique in call graph pruning task. CGPRUNER constructs a decision tree based on the 11 types of structural features (listed in Table 1) extracted from the call graphs to identify false positive edges. We run CGPRUNER using the paper’s replication package³. CGPRUNER uses a classification threshold to determine if an edge is a false positive. A higher threshold results in higher precision as it accepts only a few edges which are more likely to be the ground-truth call graph. The threshold enables a trade-off between precision and recall. We report the result of CGPRUNER at two different thresholds. The first is determined

based on the “balanced point”, where the threshold is tuned such that the average precision and recall of the call graph are equal when evaluated on the test dataset, following the procedure used by Utture et al. [51] to determine the optimal balance between precision and recall. Note that, in the CGPRUNER paper, the “balanced points” are identified using the evaluation dataset. In this paper, we selected the points using the training dataset to reduce likelihood of overfitting. The second threshold is the default threshold obtained from the replication package. Note that, for WALA, both thresholds share the same value of 0.45, so we report just one set of results. The thresholds are shown in Table 2.

- GCN [54] is a standard, off-the-shelf, graph neural network (GNN). In this task, we use it to perform edge classification. GNNs are common for machine learning on graph-structured data. As input, we pass the call graph constructed from static analysis along with 11 types of structural features described in Table 1. Then, we use the GCN, which considers information based on the nodes in the neighborhood, to predict if an edge in the call graph is a false positive.

4.2.5 Implementation details. For AUTOPRUNER, we implement the proposed approach using PyTorch library and the Python programming language. The models are trained and evaluated on two NVIDIA RTX 2080 Ti GPU with 11GB of graphics memory. For CodeBERT, we fine-tune the model with a learning rate of $1e-5$, following prior works [11, 57] and a batch size of 10 in 2 epochs. We trained the neural classifier with a learning rate of $5e-6$ with a batch size of 50 in 5 epochs.

4.3 Experimental Results

4.3.1 Effectiveness of AutoPruner. We report the comparison of our approach, AUTOPRUNER against the baselines approaches. The detailed results are shown in Table 3.

The evaluation results demonstrate that AUTOPRUNER successfully boosts the performance of the call graphs produced by WALA, Doop, and Petablox by 0.25–0.34, up to 100% improvement ($(0.68-0.34)/0.34$) in F-measure. Overall, the use of AUTOPRUNER led to gains in Precision (up to 178% improvements) which are substantially larger than the slight losses in Recall (up to just 24%).

With respect to the state-of-the-art baseline, CGPRUNER, AUTOPRUNER further improves the baseline by 13% in terms of F-measure for the call graph produced by WALA. For the call graph of Doop and Petablox, our approach improves over the optimally balanced CGPRUNER by 8% and 7%, respectively. The improvements of AUTOPRUNER over CGPRUNER in F-measure are statistically significant (p -value < 0.05) using a Wilcoxon signed-rank test.

Note that the results above, obtained from $CGPRUNER_{bal}$, is from CGPRUNER with a classification threshold carefully tuned on the testing dataset to produce the optimal balance between precision and recall. Therefore, $CGPRUNER_{bal}$ represents the optimal performance of CGPRUNER given the testing dataset. It may not always be possible to obtain the optimal threshold in practice. Despite that, we observe that AUTOPRUNER still outperforms CGPRUNER on call graphs produced by all three static analysis tools with improvements in F-measure ranging from 7%–13%.

³<https://doi.org/10.5281/zenodo.5177161>

Table 3: Comparison of the effectiveness of AUTOPRUNER with the baselines on static call graph generated by WALA, Doop, and Petablox. CGPRUNER_{bal} and CGPRUNER_{def} denotes the result of CGPRUNER at balanced point (where the precision and recall are equal on the test dataset) and default threshold (as provided in replication package), respectively. For WALA, these thresholds are the same, so we report only one set of results as CGPRUNER . The bold and underlined number denotes the best result for F-measure.

Tool	Technique	Precision	Recall	F-measure
WALA	<i>original</i>	0.24 ± 0.21	0.95 ± 0.14	0.34 ± 0.24
	<i>random</i>	0.24 ± 0.21	0.48 ± 0.07	0.27 ± 0.17
	CGPRUNER	0.66 ± 0.19	0.66 ± 0.32	0.60 ± 0.25
	GCN	0.48 ± 0.2	0.74 ± 0.37	0.54 ± 0.29
	AUTOPRUNER	0.69 ± 0.21	0.71 ± 0.19	<u>0.68 ± 0.19</u>
Doop	<i>original</i>	0.23 ± 0.21	0.92 ± 0.14	0.33 ± 0.25
	<i>random</i>	0.23 ± 0.22	0.46 ± 0.07	0.26 ± 0.17
	CGPRUNER_{bal}	0.67 ± 0.19	0.67 ± 0.30	0.61 ± 0.22
	CGPRUNER_{def}	0.72 ± 0.19	0.53 ± 0.32	0.54 ± 0.26
	GCN	0.49 ± 0.23	0.77 ± 0.31	0.55 ± 0.25
	AUTOPRUNER	0.64 ± 0.24	0.75 ± 0.15	<u>0.66 ± 0.19</u>
Petablox	<i>original</i>	0.30 ± 0.25	0.89 ± 0.14	0.40 ± 0.27
	<i>random</i>	0.3 ± 0.24	0.45 ± 0.08	0.31 ± 0.18
	CGPRUNER_{bal}	0.67 ± 0.19	0.67 ± 0.27	0.61 ± 0.20
	CGPRUNER_{def}	0.73 ± 0.19	0.52 ± 0.35	0.52 ± 0.29
	GCN	0.52 ± 0.23	0.78 ± 0.28	0.58 ± 0.23
	AUTOPRUNER	0.67 ± 0.21	0.69 ± 0.21	<u>0.65 ± 0.18</u>

When compared against CGPRUNER_{def} , which is not optimally balanced on the testing dataset and uses the threshold listed in Table 2, the improvements of AUTOPRUNER are more evident. In terms of F-measure, AUTOPRUNER outperforms CGPRUNER by 13%–25%.

Our approach performs better than the GCN baseline by 26%, 20%, and 12% in terms of F-measure for call graph of WALA, Doop, and Petablox, respectively. Interestingly, GCN underperforms CGPRUNER . Both GCN and CGPRUNER use only structural features from the call graph and differ only in the classifiers used (decision tree versus a graph neural network). This shows that for our task of call graph pruning, the more complex classifier does not outperform the simpler classifier. One possible reason for this result is that the increased complexity of the GCN causes it to overfit the training data. Overall, AUTOPRUNER outperforming both GCN and CGPRUNER suggests that the semantic features used by AUTOPRUNER have predictive power.

Answer to RQ1: Overall, AUTOPRUNER outperforms every baseline approach, including the state-of-the-art call graph pruner. The call graphs pruned by AUTOPRUNER improves over the state-of-the-art baseline by up to 13% in F1 when the baseline is optimally balanced and by up to 25% when it is not. Overall, AUTOPRUNER outperforms all baselines.

4.3.2 Effect on Client Analyses. To investigate the effect of our pruned call graph on the client analyses (i.e., static analysis tools),

we apply our call graph to two client analyses, i.e., null-pointer analysis and monomorphic call-site detection, following the experimental setup of prior work [51].

Table 4: Comparison of the effectiveness of AUTOPRUNER with the baselines and original call graph on null-pointer analysis. The bold and underlined number denotes the best result for F-measure.

Techniques	Total warnings	False Alarms Rate
<i>original</i>	8,842	73%
CGPRUNER	757	23%
AUTOPRUNER	915	<u>12%</u>

Null-pointer analysis. To investigate the impact of AUTOPRUNER in null pointer analysis, we pass the pruned call graph as input to a null pointer analysis [18], implemented in WALA. This analysis produces a set of warnings. Each warning is associated with a code location. If the call graph is less accurate, the null-pointer analysis produces more false alarms.

To evaluate AUTOPRUNER and the baseline tools, we use the same evaluation procedure as Utture et al. [51] by performing manual analysis on the reported warnings. Specifically, one author (with four years of coding experience) and one non-author (with two years of coding experience) annotator independently manually inspect warnings produced by an analysis [18] implemented in WALA. A warning is considered a “true alarm” if the author can trace the backward slice of a dereference to an instance field that was uninitialized by the end of a constructor [51]. If another exception is encountered before dereferencing the null pointer, or if the label of a warning cannot be verified in 10 minutes or is otherwise unverifiable by the authors, then the warning is considered a “false alarm”. This labelling criteria for the human annotators considers only the warnings produced by the program analysis [18], and therefore, is not a complete definition of a null pointer dereference. It is designed for an analysis that is within the cognitive ability of a human annotator to assess the call graphs produced by call graph pruners. In the cases where the two annotators disagree on the decision, we involve a third annotator, an author of the paper (with three years of coding experience), for a discussion to reach a consensus. Finally, we compute the false alarm rate of null-pointer analysis by dividing the number of false alarms by the number of warnings. The results are presented in Table 4. Furthermore, to assess the inter-rater reliability of the two annotators, we compute Cohen’s Kappa [10] and obtained a value of 0.93, which is considered as almost perfect agreement [25].

Using the call graph pruned by AUTOPRUNER, the null pointer analysis produces warnings with a false alarm rate of just 12%. Christakis and Bird [9] suggest that program analysis should aim for a false alarm rate no higher than 15-20%, which is satisfied by AUTOPRUNER. Meanwhile, both the original call graph and call graph produced by CGPRUNER resulted in false alarm rates higher than 20% (72% and 23% respectively). Our approach has a false alarm rate that is less than half of the CGPRUNER ’s false positive rate while reporting 158 more warnings. With respect to the original call graph, our approach reduces the proportion of false alarms by six times, from 73% to 12%.

Table 5: Comparison of the effectiveness of AUTOPRUNER with the baselines on monomorphic call-site detection using the call graph produced by WALA, Doop and Petablox. CGPRUNER_{bal} and CGPRUNER_{def} denotes the result of CGPRUNER at balanced point (where the precision and recall are equal) and default threshold (as provided in the replication package), respectively. For WALA, these thresholds are the same, so we report only one set of results for CGPRUNER. The bold and underlined number denotes the best result for F-measure.

Tool	Techniques	Precision	Recall	F-measure
WALA	<i>original</i>	0.52 ± 0.23	0.93 ± 0.15	0.64 ± 0.21
	CGPRUNER	0.68 ± 0.18	0.68 ± 0.32	0.62 ± 0.25
	AUTOPRUNER	0.71 ± 0.22	0.72 ± 0.19	0.69 ± 0.19
Doop	<i>original</i>	0.51 ± 0.24	0.92 ± 0.14	0.63 ± 0.22
	CGPRUNER _{bal}	0.68 ± 0.19	0.71 ± 0.31	0.63 ± 0.23
	CGPRUNER _{def}	0.68 ± 0.21	0.56 ± 0.37	0.53 ± 0.30
	AUTOPRUNER	0.66 ± 0.24	0.78 ± 0.16	0.69 ± 0.20
Petablox	<i>original</i>	0.52 ± 0.23	0.9 ± 0.15	0.63 ± 0.21
	CGPRUNER _{bal}	0.68 ± 0.19	0.72 ± 0.28	0.63 ± 0.21
	CGPRUNER _{def}	0.73 ± 0.19	0.58 ± 0.34	0.56 ± 0.26
	AUTOPRUNER	0.69 ± 0.21	0.75 ± 0.20	0.68 ± 0.19

Monomorphic call-site detection. In the task of monomorphic call site detection, the call graph is used to identify call sites where there is only one concrete call at a code location. As shown in Table 5, for the call graph constructed by WALA, AUTOPRUNER outperforms the original call graph and CGPRUNER in F-measure. AUTOPRUNER outperforms CGPRUNER by 11% in F-measure, with improvements in both precision and recall.

We observe similar performances on the call graph of Doop and Petablox. On Doop’s call graph, AUTOPRUNER outperforms CGPRUNER by 30% in F-measure. Compared to the original graph, AUTOPRUNER improves in F-measure by 10%. Similarly, on the call graph produced by Petablox, AUTOPRUNER improves over both CGPRUNER and the original call graph by 8% in terms of F-measure.

Answer to RQ2: The call graph produced by AUTOPRUNER leads to improvements in both null pointer analysis and monomorphic call site detection. Based on the call graph from WALA, AUTOPRUNER decreases the false alarm rate from null pointer analysis by 11%. On monomorphic call site detection, AUTOPRUNER improves over the state-of-the-art call graph pruner by 11% in F-measure.

4.3.3 Ablation Study. In answer this question, we investigate two different ablation studies:

- Semantic versus Structure
- Caller versus Callee function

Semantic vs. Structure. In this experiment, we evaluate the relative contribution of AUTOPRUNER’s semantic versus structural features for call graph pruning. Table 6 shows the results of our experiments. AUTOPRUNER_{sem} refers to AUTOPRUNER using only the semantic features extracted from the source code, and AUTOPRUNER_{struct},

Table 6: Comparison of the effectiveness of the semantic features and the structural features. The bold and underline number denotes the best result for F-measure.

Techniques	Precision	Recall	F-measure
AUTOPRUNER _{struct}	0.58	0.75	0.62
AUTOPRUNER _{sem}	0.67	0.71	0.66
AUTOPRUNER	0.69	0.71	0.68

refers to the AUTOPRUNER using only the structural features. Using only the semantic features, AUTOPRUNER_{sem} outperforms AUTOPRUNER_{struct} in F-measure by 6%. By using with both types of features, AUTOPRUNER outperforms AUTOPRUNER_{sem} by 3% in F-measure. The decreases in F-measure when either the semantic features or structural features are removed are statistically significant (p-value < 0.05). This indicates that both types of features are important, but relatively larger decrease in F-measure when removing the semantic features indicates that the semantic features are more important compared to the structural features.

Overall, when AUTOPRUNER uses only one type of feature, AUTOPRUNER has a lower F-measure. This suggests that both semantic and structural features are important for AUTOPRUNER to perform effectively.

Table 7: Comparison of the effectiveness of the caller features and the callee features. The bold and underline number denotes the best result for F-measure.

Techniques	Precision	Recall	F-measure
AUTOPRUNER _{caller}	0.69	0.60	0.58
AUTOPRUNER _{callee}	0.68	0.66	0.60
AUTOPRUNER	0.69	0.71	0.68

Caller vs. Callee. Next, we assess the relative importance of the features extracted from the caller and callee functions. We evaluate the performances of AUTOPRUNER when only considering source code from either caller and callee and compare them with AUTOPRUNER’s. To perform this study, we fine-tuned CodeBERT with only either the source code of the caller or the callee function.

AUTOPRUNER_{caller} refers to AUTOPRUNER using only the source code from the caller function, and AUTOPRUNER_{callee} refers to AUTOPRUNER using only the source code from the callee function. AUTOPRUNER outperforms AUTOPRUNER_{caller} and AUTOPRUNER_{callee} by up to 17%. The decreases in F-measure are statistically significant (p-value < 0.05). This suggests that the semantic features extracted from the source code of both the caller and callee functions are crucial for the performance of AUTOPRUNER.

Answer to RQ3: Our ablation study shows that all features contribute to the effectiveness of AUTOPRUNER. The semantic features are more important than the structural features, while both the caller and callee functions are essential.

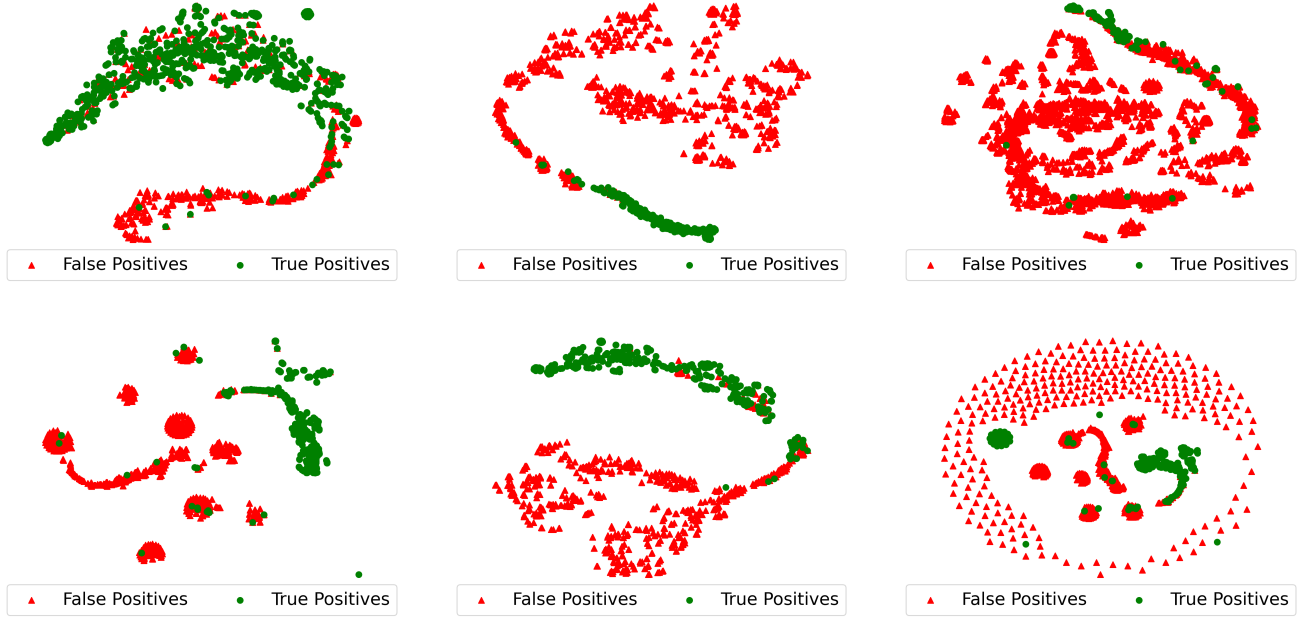


Figure 3: The visualization of semantic features produced by *CodeBERT* in 6 sample programs. The green circles and red triangles represent true positive and false positive edges respectively.

5 DISCUSSION

5.1 Qualitative Analysis

In this section, we perform a qualitative analysis of *AUTOPRUNER*. We have seen that *AUTOPRUNER* consistently outperforms the state-of-the-art approach, and our ablation study indicates that the semantic features extracted by *CodeBERT* are essential for the effective performance of *AUTOPRUNER*. Here, our goal is to investigate if the pre-trained transformer model of code, i.e., *CodeBERT*, is able to separate true positive edges from false positives in the call graph. To this end, we use *t-SNE* [52], an unsupervised method for visualization, to visualize the semantic features from the call graphs of 6 programs in two-dimensional space. If the semantic features have predictive power, we would expect the call graph edges, which are false positives, to be separated from the true positives.

Figure 3 presents the visualizations, where the green and red points are features of true-positive and false-positive edges, respectively. Indeed, we observe that the majority of the green points are clustered and are located relatively close to one another. Furthermore, the clusters of green points are also typically far away from the majority of the red points. The observation suggests that *CodeBERT* was able to be trained to extract semantic features from the caller and callee function such that the true positives can be separated from the false positives in the vector space. This validates our use of a fine-tuned *CodeBERT* model for extracting semantic features, as the model demonstrates a remarkable ability to distinguish between true and false positives.

Still, a small but significant proportion of red and green points are located close to one another, indicating that using only semantic features is not enough to separate these edges. This suggests that other features (e.g., structural features) should be used to improve the model’s classifier ability. Indeed, as shown in Section 4.3.3, the structural features are complementary to the semantic features. The addition of structural features increases the Precision of *AUTOPRUNER* by 3% while keeping the same Recall, leading to 3% improvement on F-Measure. Nevertheless, we acknowledge the modest contribution of the structural features which may be from how the structural and semantic features are combined. Currently, after they are independently extracted, they are combined with a small feed-forward neural network (FFNN). As a result, the FFNN may fail to capture more complex interactions between the semantic and structural features.

5.2 Efficiency

In this section, we investigate the efficiency of *AUTOPRUNER*. For pre-training model, *AUTOPRUNER* uses an existing pretrained model, *CodeBERT*⁴. For the offline fine-tuning (Section 3.1) and training (Section 3.2) phase wherein both the *CodeBERT* and the binary classifier needs to be finetuned and trained only once, *AUTOPRUNER* takes around 36 hours. For the inference phase which is integrated in downstream applications, *AUTOPRUNER* takes around 0.04 second on average to predict a label for an edge.

⁴<https://github.com/microsoft/CodeBERT>

5.3 Threats to validity

5.3.1 External validity. Threats to external validity concern the generalizability of our findings. Our experiments are performed on the same dataset as prior work [51], constructed from the NJR-1 benchmark [36]. This may be a threat to external validity since AUTOPRUNER may not generalize beyond the programs outside the NJR-1 dataset. However, this threat is minimal as the dataset consists of a large number of data points, and the NJR-1 benchmark is large and was carefully constructed to ensure their diversity [36].

5.3.2 Internal validity. Threats to internal validity refer to possible errors in our experiments. In this study, following the experimental procedure of prior work [51], we perform a manual inspection of null-pointer analysis, which may introduce human error. To minimize the risk, we asked one author of this paper and a non-author to independently inspect and label the reported warnings. We have measured the inter-rater reliability, obtaining Cohen’s kappa of 0.93, which can be interpreted as almost perfect agreement [25]. Therefore, we believe that there are minimal threats from this issue.

5.3.3 Construct validity. Threats to construct validity relate to the suitability of our evaluation. To minimize risks to construct validity, we have used the same experimental setup as a previous study [51], including the same dataset and ground-truth labels. Some bias could be introduced in the construction of the ground truth. Manual labeling requires extensive human effort, which limits the scale of the experiments. Hence, we use the same automated labeling procedure from prior work [51] that runs test cases to identify ground truth edges. The imperfect code coverage of the test cases may introduce bias. However, the code coverage in our experiments (68%) is higher than the code coverage of real-world programs (less than 40%) reported in prior studies [24].

Another threat to construct validity is the definition of null pointer analysis of warnings produced by [18]. For analyzing static analysis warnings, prior works often employ a human study with several annotators [14, 49]. This human study is expensive and the task given to the annotators must be within the cognitive ability of humans. To make the annotation task tractable to humans and reduce its cost, we use the true-alarm identification task definition used in the prior work [51].

6 RELATED WORK

In Section 2, we have discussed the studies related to call graph pruning and CodeBERT. Here, we discuss other related studies.

Call graph construction has been widely studied. As our approach does not use runtime information, it falls into the class of static approaches [32, 41, 46, 51] for constructing call graphs. Approaches that use dynamic analysis [15, 55] result in fewer false positives but are less scalable.

Some recent studies on call graph construction have focused on dynamic languages. Salis et al. [45] and Nielson et al. [35] present approaches for constructing call graphs of Python programs and Javascript programs. Unlike language-specific techniques, AutoPruner can be used to improve call graphs of any language.

There are many client analyses using call graphs. Recently, call graphs have been used for practical applications, including scanning applications for vulnerable library usage [35], generating exploits of

vulnerabilities [19], and impact analysis [15]. We have explored two classic client analyses, null pointer analysis and monomorphic call site detection, to validate that the improvements from pruning the call graph lead to further improvements in practical applications.

Apart from applying CodeBERT to call graph construction, researchers have proposed other applications of deep learning models on source code. Other researchers had success using deep learning models for type inference [2, 16, 37, 38], code completion [17, 40, 47], code clone detection [44], program repair [8, 31], fault localization [28, 33], among other analyses on source code [6, 26, 29, 34]. Our work is similar as AutoPruner successfully applies deep learning techniques on source code analysis, but is unique as previous studies have not previously applied deep learning to call graph analysis.

7 CONCLUSION AND FUTURE WORK

We propose AUTOPRUNER, a novel call graph pruner that leverages both structural and semantic features. AUTOPRUNER employs CodeBERT to extract semantic features from both the caller and callee function associated with each edge in the call graph. Our empirical evaluation shows that AUTOPRUNER outperforms multiple baselines, including the state-of-the-art approach that uses only structural features. The improvements from AUTOPRUNER also lead to tangible improvements on downstream applications. In particular, the proportion of false alarms reported by a baseline null pointer analysis is halved, decreasing from 23% to just 12%.

Our ablation study shows that the semantic features complement the structural features. Moreover, our qualitative analysis reveals that the semantic features extracted by CodeBERT effectively separate true and false positive edges.

In the future, we will evaluate AUTOPRUNER on call graphs constructed using additional static analysis tools and for other programming languages, as well as assess the impact on call graph pruning on other client analyses. We will also explore more ways to improve AUTOPRUNER, such as jointly extracting semantic and structural features, and providing more contextual information by incorporating the k -hop callers (e.g. the caller of the caller) of each function call to enrich the semantic features. Another interesting direction is to use AUTOPRUNER for proposing edges in the call graph missing due to an unsoundness program analysis.

Data availability. AutoPruner’s dataset and implementation are publicly available at <https://github.com/soarsmu/AutoPruner/>.

ACKNOWLEDGEMENT

This project is supported by the National Research Foundation, Singapore and National University of Singapore through its National Satellite of Excellence in Trustworthy Software Systems (NSOE-TSS) office under the Trustworthy Computing for Secure Smart Nation Grant (TCSSNG) award no. NSOE-TSS2020-02. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and National University of Singapore (including its National Satellite of Excellence in Trustworthy Software Systems (NSOE-TSS) office).

Xuan-Bach D. Le is supported by the Australian Government through the Australian Research Council’s Discovery Early Career Researcher Award, project number DE220101057.

REFERENCES

- [1] Karim Ali and Ondřej Lhoták. 2012. Application-only call graph construction. In *European Conference on Object-Oriented Programming*. Springer, 688–712.
- [2] Miltiadis Allamanis, Earl T Barr, Soline Ducousso, and Zheng Gao. 2020. Typilus: Neural type hints. In *Proceedings of the 41st acm sigplan conference on programming language design and implementation*. 91–105.
- [3] Leonhard Applis, Annibale Panichella, and Arie van Deursen. 2021. Assessing Robustness of ML-Based Program Analysis Tools using Metamorphic Program Transformations. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1377–1381.
- [4] Martin Bravenboer and Yannis Smaragdakis. 2009. Strictly declarative specification of sophisticated points-to analyses. In *Proceedings of the 24th ACM SIGPLAN conference on Object oriented programming systems languages and applications*. 243–262.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [6] Nghi DQ Bui, Yijun Yu, and Lingxiao Jiang. 2021. Infercode: Self-supervised learning of code representations by predicting subtrees. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1186–1197.
- [7] David Callahan, Alan Carle, Mary W. Hall, and Ken Kennedy. 1990. Constructing the procedure call multigraph. *IEEE Transactions on Software Engineering* 16, 4 (1990), 483–487.
- [8] Zimin Chen, Steve Kommrusch, Michele Tufano, Louis-Noël Pouchet, Denys Poshyvanyk, and Martin Monperrus. 2019. Sequencer: Sequence-to-sequence learning for end-to-end program repair. *IEEE Transactions on Software Engineering* 47, 9 (2019), 1943–1959.
- [9] Maria Christakis and Christian Bird. 2016. What developers want and need from program analysis: an empirical study. In *Proceedings of the 31st IEEE/ACM international conference on automated software engineering*. 332–343.
- [10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [11] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1536–1547.
- [12] Stephen Fink and Julian Dolby. 2012. WALA—The TJ Watson Libraries for Analysis.
- [13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.
- [14] Andrew Habib and Michael Pradel. 2018. How many of all bugs do we find? a study of static bug detectors. In *2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 317–328.
- [15] Joseph Hejderup, Arie van Deursen, and Georgios Gousios. 2018. Software ecosystem call graph for dependency management. In *2018 IEEE/ACM 40th International Conference on Software Engineering: New Ideas and Emerging Technologies Results (ICSE-NIER)*. IEEE, 101–104.
- [16] Vincent J Hellendoorn, Christian Bird, Earl T Barr, and Miltiadis Allamanis. 2018. Deep learning type inference. In *Proceedings of the 2018 26th acm joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 152–162.
- [17] Abram Hindle, Earl T Barr, Mark Gabel, Zhendong Su, and Premkumar Devanbu. 2016. On the naturalness of software. *Commun. ACM* 59, 5 (2016), 122–131.
- [18] Laurent Hubert, Thomas Jensen, and David Pichardie. 2008. Semantic foundations and inference of non-null annotations. In *International Conference on Formal Methods for Open Object-Based Distributed Systems*. Springer, 132–149.
- [19] Emanuele Iannone, Dario Di Nucci, Antonino Sabetta, and Andrea De Lucia. 2021. Toward Automated Exploit Generation for Known Vulnerabilities in Open-Source Libraries. In *2021 IEEE/ACM 29th International Conference on Program Comprehension (ICPC)*. IEEE, 396–400.
- [20] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. 2013. Why don't software developers use static analysis tools to find bugs?. In *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, 672–681.
- [21] Milod Kazerounian, Jeffrey S Foster, and Bonan Min. 2021. SimTyper: sound type inference for Ruby using type equality prediction. *Proceedings of the ACM on Programming Languages* 5, OOPSLA (2021), 1–27.
- [22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
- [23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
- [24] Pavneet Singh Kochhar, David Lo, Julia Lawall, and Nachiappan Nagappan. 2017. Code coverage and postrelease defects: A large-scale study on open source projects. *IEEE Transactions on Reliability* 66, 4 (2017), 1213–1228.
- [25] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
- [26] Alexander LeClair, Siyuan Jiang, and Collin McMillan. 2019. A neural model for generating natural language summaries of program subroutines. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 795–806.
- [27] Yue Li, Tian Tan, Anders Møller, and Yannis Smaragdakis. 2018. Scalability-first pointer analysis with self-tuning context-sensitivity. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 129–140.
- [28] Yiling Lou, Qihao Zhu, Jinhao Dong, Xia Li, Zeyu Sun, Dan Hao, Lu Zhang, and Lingming Zhang. 2021. Boosting coverage-based fault localization via graph-based representation learning. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 664–676.
- [29] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664* (2021).
- [30] Ravi Mangal, Xin Zhang, Aditya V Nori, and Mayur Naik. 2015. A user-guided approach to program analysis. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. 462–473.
- [31] Ehsan Mashhadi and Hadi Hemmati. 2021. Applying codebert for automated program repair of java simple bugs. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. IEEE, 505–509.
- [32] Gail C Murphy, David Notkin, William G Griswold, and Erica S Lan. 1998. An empirical study of static call graph extractors. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 7, 2 (1998), 158–191.
- [33] Thanh-Dat Nguyen, Thanh Le-Cong, Duc-Ming Luong, Van-Hai Duong, Xuan Bach Le Dinh, David Lo, and Thang Huynh-Quyet. 2022. FFL: Fine grained Fault Localization for Student Programs via Syntactic and Semantic Reasoning. In *Proceedings of the 38th IEEE International Conference on Software Maintenance and Evolution*. IEEE.
- [34] Giang Nguyen-Truong, Thanh Le-Cong, Hong Jin Kang, Xuan-Bach Dinh Le, and David Lo. 2022. VulCurator: A Vulnerability-Fixing Commit Detector. In *Proceedings of the 30th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM.
- [35] Benjamin Barslev Nielsen, Martin Toldam Torp, and Anders Møller. 2021. Modular call graph construction for security scanning of Node.js applications. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 29–41.
- [36] Jens Palsberg and Cristina V Lopes. 2018. NJR: A normalized Java resource. In *Companion Proceedings for the ISSTA/ECCOP 2018 Workshops*. 100–106.
- [37] Yun Peng, Cuiyun Gao, Zongjie Li, Bovei Gao, David Lo, Qirun Zhang, and Michael Lyu. 2022. Static inference meets deep learning: a hybrid type inference approach for python. In *Proceedings of the 44th International Conference on Software Engineering*. 2019–2030.
- [38] Michael Pradel, Georgios Gousios, Jason Liu, and Satish Chandra. 2020. Type-writer: Neural type prediction with search-based validation. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 209–220.
- [39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [40] Veselin Raychev, Martin Vechev, and Eran Yahav. 2014. Code completion with statistical language models. In *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 419–428.
- [41] Michael Reif, Florian Kübler, Michael Eichberg, Dominik Helm, and Mira Mezini. 2019. Judge: Identifying, understanding, and evaluating sources of unsoundness in call graphs. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 251–261.
- [42] Henry Gordon Rice. 1953. Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical society* 74, 2 (1953), 358–366.
- [43] Barbara G Ryder. 1979. Constructing the call graph of a program. *IEEE Transactions on Software Engineering* 3 (1979), 216–226.
- [44] Vaibhav Saini, Farima Farmahinifarahani, Yadong Lu, Pierre Baldi, and Cristina V Lopes. 2018. Oreo: Detection of clones in the twilight zone. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 354–365.
- [45] Vitalis Salis, Thodoris Sotiropoulos, Panos Louridas, Diomidis Spinellis, and Dimitris Mitropoulos. 2021. Pygc: Practical call graph generation in python. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1646–1657.
- [46] Li Sui, Jens Dietrich, Amjed Tahir, and George Fourtounis. 2020. On the recall of static call graph construction in practice. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*. IEEE, 1049–1060.
- [47] Alexey Svyatkovskiy, Ying Zhao, Shengyu Fu, and Neel Sundaresan. 2019. Pythia: AI-assisted code completion system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2727–2735.

- [48] Tian Tan, Yue Li, and Jingling Xue. 2016. Making k-object-sensitive pointer analysis more precise with still k-limiting. In *International Static Analysis Symposium*. Springer, 489–510.
- [49] David A Tomassi and Cindy Rubio-González. 2021. On the Real-World Effectiveness of Static Bug Detectors at Finding Null Pointer Exceptions. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 292–303.
- [50] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 242–264.
- [51] Akshay Utture, Shuyang Liu, Christian Gram Kalhauge, and Jens Palsberg. 2022. Striking a Balance: Pruning False-Positives from Static Call Graphs. *The 44rd IEEE/ACM International Conference on Software Engineering (ICSE 2022)* (2022).
- [52] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [54] Max Welling and Thomas N Kipf. 2016. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*.
- [55] Tao Xie and David Notkin. 2002. An empirical study of java dynamic call graph extractors. *University of Washington CSE Technical Report* (2002), 02–12.
- [56] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
- [57] Jiayuan Zhou, Michael Pacheco, Zhiyuan Wan, Xin Xia, David Lo, Yuan Wang, and Ahmed E Hassan. 2021. Finding A Needle in a Haystack: Automated Mining of Silent Vulnerability Fixes. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 705–716.
- [58] Xin Zhou, DongGyun Han, and David Lo. 2021. Assessing Generalizability of CodeBERT. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 425–436.
- [59] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.
- [60] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. <https://aclanthology.org/2021.ccl-1.108>