

ASTRO: An AST-Assisted Approach for Generalizable Neural Clone Detection

Yifan Zhang
yifan.zhang.2@vanderbilt.edu
Vanderbilt University
Nashville, TN, USA

Junwen Yang
junweny@fb.com
Meta, Inc.
Menlo Park, CA, USA

Haoyu Dong
hd108@duke.edu
Duke University
Durham, NC, USA

Qingchen Wang
qcwang@hku.hk
University of Hong Kong
Hong Kong SAR

Huajie Shao
hshao@wm.edu
College of William & Mary
Williamsburg, VA, USA

Kevin Leach
kevin.leach@vanderbilt.edu
Vanderbilt University
Nashville, TN, USA

Yu Huang
yu.huang@vanderbilt.edu
Vanderbilt University
Nashville, TN, USA

ABSTRACT

Neural clone detection has attracted the attention of software engineering researchers and practitioners. However, most neural clone detection methods do not generalize beyond the scope of clones that appear in the training dataset. This results in poor model performance, especially in terms of model recall. In this paper, we present an **Abstract Syntax Tree (AST)** assisted approach for **geneRALizable neural clone detectiOn**, or **ASTRO**, a framework for finding clones in codebases reflecting industry practices. We present three main components: (1) an AST-inspired representation for source code that leverages program structure and semantics, (2) a global graph representation that captures the context of an AST among a corpus of programs, and (3) a graph embedding for programs that, in combination with extant large-scale language models, improves state-of-the-art code clone detection. Our experimental results show that ASTRO improves state-of-the-art neural clone detection approaches in both recall and F-1 scores.

ACM Reference Format:

Yifan Zhang, Junwen Yang, Haoyu Dong, Qingchen Wang, Huajie Shao, Kevin Leach, and Yu Huang. 2022. ASTRO: An AST-Assisted Approach for Generalizable Neural Clone Detection. In *Proceedings of Automated Software Engineering Industrial Showcase (ASE'22)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Code *Clones* refer to syntactically or functionally similar code fragments within or between software systems [18]. Having clones in industrial codebases can substantially increase maintenance costs [10]. For example, 22.3% of operating systems' defects were introduced by code cloning [27]. Moreover, clones can also lead to intellectual property violations and security problems [4]. With the increasing scale of modern software, clone detection (CD) techniques have been developed to automatically detect clones within corpora of code. Traditional CD approaches aim to find similar code patterns

by leveraging abstract representations, including tokens [13, 21], Abstract Syntax Trees (AST) [3, 9], Program Dependency Graphs (PDG) [28], and other intermediate representations. However, these methods frequently focus on one representation but ignore others, leading to a failure to generalize across diverse software corpora.

With the rise of neural networks, CD techniques have leveraged neural architectures that learn program semantics and structure [14, 22]. For example, White et al. [24] learns both lexical and syntactic embeddings by sequence modeling, Zhang et al. [26] incorporates code structural information within a Long Short Term Memory (LSTM) to improve detection performance, and Yu et al. [25] applies machine learning to this task. These neural CD approaches are based on learning patterns of semantics in the dataset and follow common practices in deep learning. While these models achieve high precision, their applications are restricted to the domain of semantics present in the programs that comprise the training data. That is, when code fragments from other domains are placed as input to such models, the detection performance as measured by recall is close to zero. This inability to generalize across software domains can be problematic due to the large scale of industrial software codebases.

In this paper, we propose **ASTRO**, a novel framework that combines neural network architectures, Abstract Syntax Trees, and context-aware program structure to quickly and accurately detect code clones. Our approach is based on two key insights derived from domain knowledge in software engineering: (1) identical code snippets share identical AST structures. Even though large programs can have large ASTs, there is a limited number of node types and topologies of ASTs for a given programming language. This insight allows us to capture the key structural characteristics of a program and detect code clones that only differ on detailed design (i.e., variable names, adding or deleting statements, etc.). (2) Unseen code snippets can be represented and retrieved with a combination of previously-seen (sub-) ASTs. With this insight, we build a global graph with training data (i.e., programs) that captures structural and co-occurrence information of nodes in source ASTs (see Section 3.2). Previous work such as ASTNN [26] have also considered neural networks and graph embeddings of ASTs as monolithic, end-to-end systems. However, our approach can be used as a plug-in to existing pretrained models (such as CodeBERT, RoBERTa, and ASTNN) to enhance the performance and generalizability of code clone detection.

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.

ASE'22, 2022, Ann Arbor, Michigan

© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXX.XXXXXXX>

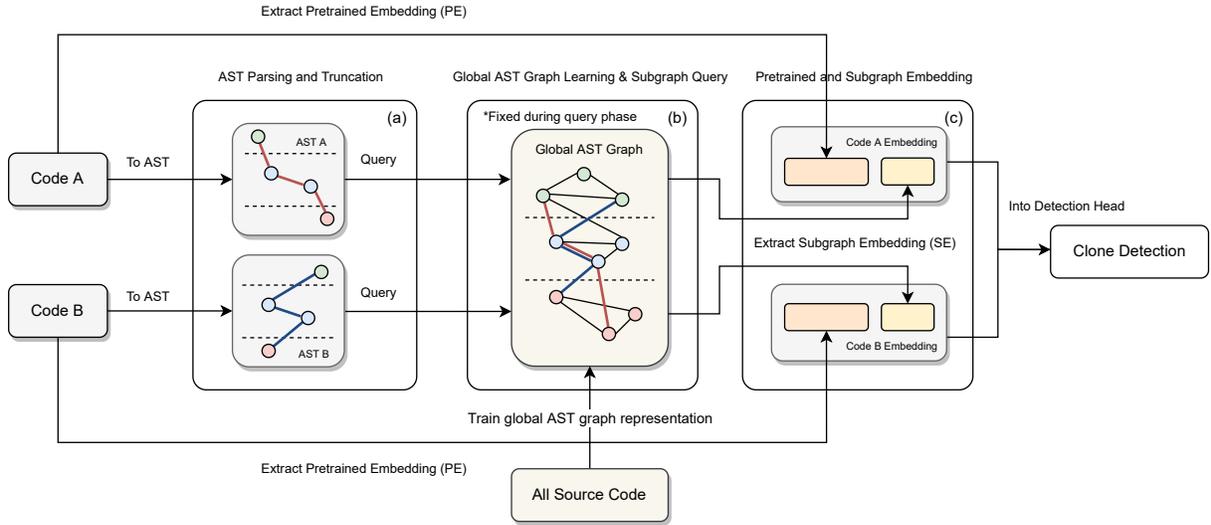


Figure 1: Diagram of ASTRO. It consists of three main components: (a) AST parsing and truncation 3.1, (b) global AST graph learning and matching 3.2 and (c) pretrained and subgraph embedding extraction 3.3. The global AST graph in (b) is built from the training data. In practice, Code A and Code B would be snippets of source that are not part of the training set.

At a high level, ASTRO consists of three components, illustrated in Figure 1. First, we **parse all training programs into ASTs with k -hierarchy truncation** and build a **global AST graph** accordingly. Then, each pair of code snippets to be considered for clone detection are **queried from the global AST graph with their truncated AST representations**. Lastly, we introduce a novel strategy to generate **merged subtree embeddings** for each code pair as plug-in to their embeddings extracted from a pretrained model, and input the concatenated embedding to our non-linear detection head to improve the generalizability of clone detection. We elaborate on the design of ASTRO in Section 3.

We evaluate ASTRO using Big Clone Bench [20] and compare against typical CD methods [9] and more recent neural CD methods [19, 24]. Our results show that, while ASTRO sacrifices precision by less than 0.19, it substantially improves the recall by 0.85 and F-1 score by 0.80. This indicates that ASTRO is capable of more readily generalizing clone detection across input programs across software domains. The main contributions of this paper are:

- ASTRO, a graph-based and AST-guided neural network model for code clone detection.
- A novel graph-based approach to capture inherent information of code snippets for generalization.
- A novel approach to leverage domain knowledge in SE to improve the large-scale pretrained models in AI.
- A systematic evaluation of ASTRO on neural clone detection and pretrained model enhancement. ASTRO improves the recall and F1 score of neural CD methods, and enhances the generalizability of large-scale pretrained models.

The rest of the paper is structured as follows: Section 2 highlights previous related work, Section 3 describes the technical approach behind the ASTRO framework, Section 4 presents our evaluation and discussion for ASTRO, and Section 5 concludes the paper.

2 BACKGROUND AND RELATED WORK

Code clones are widespread and prevalent since developers frequently reuse code via copy-paste to reduce programming effort. Previous studies indicate that 12% of Linux code was identified as code clones [12], and 19% of the X Window system functionality entails code clones [2].

Unfortunately, code clones increase the likelihood of errors and software defects. Many studies have indicated serious negative impacts of code clones in software evolution, including hidden bug propagation [17], unintentional inconsistencies [7], and high instability [16]. Substantial research has been conducted to detect code clones and related defects.

Traditional code clone detection. Before the renaissance of neural networks, code clone detection was categorized in three ways: string-based, in which substrings of program source code were directly compared [2], token-based, in which lexical program token sequences were directly compared [11], and parser-based, in which graph structures like ASTs and PDGs were directly compared [3, 9]. However, these methods often overfit or miss complex cases in which code may be semantically similar but structurally different (e.g., inlined functions, renamed variables, or nested loops in different order).

Neural code clone detection More recently, several works based on neural networks have been explored for code clone detection, which have focused on learning abstract representations of code such as ASTs [26], PDGs [28], and CFGs [5]. Although these methods can achieve high precision (i.e., given a source code input, they can identify a low number of clones), they fail to generalize when subtle changes exist in code clones, resulting in poor recall (i.e., given a source code input, these models cannot identify all clones of that input).

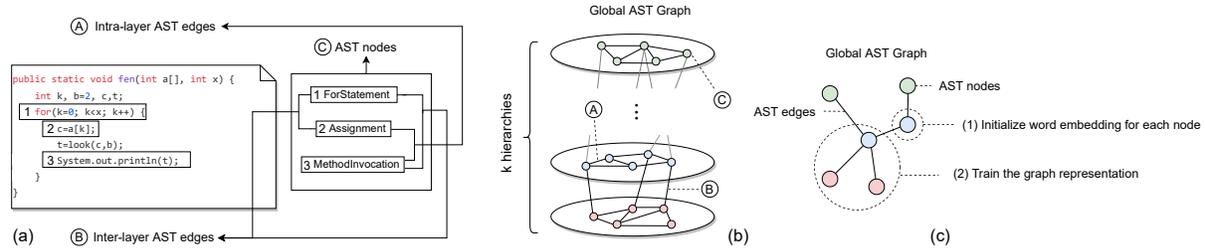


Figure 2: Illustration of the global AST graph construction and learning. Subfigure (a) shows an example program and AST, Subfigure (b) shows how inter- and intra-hierarchy edges are modeled in the global graph, and Subfigure (c) shows how each node embedding is initialized and trained to form the global graph.

3 APPROACH

ASTRO consists of three main components, shown in Figure 1. Our task is to identify whether two input code snippets are clones, thus it is a supervised binary classification task. Given two source code snippets as input, we first parse them into ASTs, then derive an embedding based upon fixed-depth walks of each subtree in each AST (Section 3.1). Then, we generate a *global graph* that captures the relationships that exist between subtrees of an AST (e.g., methods are defined in classes, thus method subtrees are related to class subtrees). Our global graph construction allows for the creation of a merged embedding that captures both structure (by matching ASTs and subtrees between snippets) and semantics (by capturing the context in which various subtrees appear) (Section 3.2). Finally, we use a non-linear detection head to determine whether the two input snippets are clones (Section 3.3). We discuss each step below.

3.1 AST Parsing and Truncation

Each pair of code for clone detection is first parsed into ASTs. In this paper, we consider Java programs for our evaluation, we parse using the JavaLang parser¹. JavaLang’s abstract grammar contains 19 node types. As ASTs are recursive structures, they can be an unlimited height depending on the size and complexity of a given program. However, we hypothesize that very deep nodes in the AST are less important for conveying generalizable program semantics than more shallow nodes. Thus, we derive an embedding of an input program’s AST by truncating the tree to a fixed height k . We refer to each layer of sibling nodes as a *hierarchy* of nodes kept within the truncated AST. In our experiments, we set $k = 5$, meaning we keep all nodes of height 5 or less from the input AST.

3.2 Global AST Graph Learning and Subgraph Query

To address the challenge of generalizability, we construct a *global AST graph* using truncated ASTs obtained from the previous step (Section 3.1). We leverage two pieces of information from all programs in a training corpus to build the global AST graph: (1) the types of nodes in the ASTs, and (2) edges between nodes within each AST. Both are available at each level of each AST. An indicative example is shown in Figure 2. The input program is parsed

into an AST, from which k hierarchies of nodes are kept (Figure 2 (a)). When quantifying the co-occurrences of different nodes within an AST, we model *inter-hierarchy* edges (e.g., ForStatement \rightarrow Assignment) and *intra-hierarchy* edges (e.g., Assignment \rightarrow MethodInvocation) in the construction of the global AST graph (Figure 2 (b)). Each AST node is represented by an embedding (256 dimensions in our experiments), which reflects an average of embeddings of each child (Figure 2 (c)).

To build the global AST graph, we apply unsupervised graph representation learning with the GraphSAGE algorithm [8]. GraphSAGE is a graph representation learning method that learns the embeddings of nodes in the graph based on graph structure and node features. Before training the global graph representation, we use Flair word embeddings [1] to initialize the word embedding for each name of each node type (e.g., “ForControl”) in the AST with a dimension of 256 (adopted from GraphSAGE). For sequence aggregation in GraphSAGE, we apply the default mean aggregator and obtain a 256-dimension embedding vector for every node in the global graph. Due to the large scale of edge modeling in the global AST graph, we set the sampling ratio of all edges to 0.1% to mitigate the risk of the over smoothing problem.

With the global AST graph trained by the training corpus of programs, to compare whether two code snippets are clones, we make queries of their ASTs against the global graph to obtain their corresponding node representations (i.e., a subgraph from the global graph), as shown in Figure 1 (a) and (b). In the query process, we finalize the representation for the subgraph (corresponding to the given AST) using all its node embeddings based on its tree structure. Specifically, starting from the bottom of the AST, we take the average of all embeddings of each node’s children, then repeat the process until we reach the top of the tree. The final subgraph embedding of the AST is a single 256-dimensional vector that reflects an average of all of the nodes in the AST.

3.3 Pretrained and Subgraph Embedding Extraction

Finally, we concatenate the subgraph embedding (Section 3.2) with a pretrained embedding (of the source code tokens) as our final ASTRO embedding to be used for clone detection, shown in Figure 1 (c). In ASTRO, the pretrained embedding can be adopted from any state-of-the-art language model, such as RoBERTa [15] or CodeBERT [6]. Since the AST representation only reflects program syntax and

¹JavaLang is available at <https://github.com/c2nes/JavaLang>

structure, we further introduce program semantics by adopting pretrained code models to our detection framework. For example, by adopting RoBERTa, the final code embedding concatenates a 768-dimension pretrained embedding with our 256-dimension subgraph embedding. This strategy improves the overall representation of the input code snippets for the clone detection task.

To make a final prediction, we integrate a non-linear detection head for clone detection in ASTRO, following common practice. Specifically, for our experiments, we adopt the detection head in CodeBERT [6] and simplify the structure to three tanh-activated non-linear layers with 512, 256, and 128 hidden nodes, respectively, to test the effectiveness of our ASTRO plug-in. We use a learning rate of 1e-3, dropout rate of 0.1, batch size of 512, and 50 epochs during training.

4 EMPIRICAL EVALUATION

4.1 Dataset

We evaluate ASTRO on Big Clone Bench (BCB) [20], a widely-used dataset for clone detection. The training, validation, and testing splits follow the state-of-the-art work: 901,028 training code pairs, and 415,416 validation/testing code pairs.

4.2 Evaluation Results

We evaluate the performance of ASTRO with three metrics: precision, recall, and F1 score. We compare ASTRO with multiple traditional and neural CD models. As shown in Table 1, in the clone classification task, Deckard [9], DLC [24], and SourcererCC [19] models have relatively higher precision, but very low total recall, indicating their ineffectiveness in revealing structurally similar but semantically diverse clones. In contrast, ASTRO sacrifices a trivial amount of precision (- 0.19) compared with DLC, but achieves much higher recall (+ 0.85) and F1 (+ 0.80), demonstrating effectiveness in revealing structurally similar clones.

Table 1: Quantitative evaluation of ASTRO and other neural clone detection methods using precision, recall and F1 score. Results of Deckard, DLC and SourcererCC are from [23].

Approaches	Precision	Recall	F1
Deckard [9]	0.93	0.02	0.03
DLC [24]	0.95	0.01	0.01
SourcererCC [19]	0.88	0.02	0.03
ASTRO (Ours)	0.76	0.86	0.81

4.3 Ablation Studies

In this section, we study the contributions of each proposed module in the subgraph embedding of ASTRO as input into detection head without concatenating the pretrained embedding, denoted as ASTRO (Graph). Experimental results of three ablation studies are shown in Table 2.

In the first row, ASTRO (Graph) indicates the use of subgraph embeddings alone, without any additional concatenation to other embeddings. In (A), we remove the upward merging process of AST hierarchies and use average embeddings of all AST nodes. It shows

Table 2: Ablation studies of ASTRO (Graph). (A) removes embedding merging in the subgraph, (B) removes edge sampling before training global graph representation, and (C) removes the whole graph learning and uses word embeddings.

Approaches	Precision	Recall	F1	Recall Diff
ASTRO (Graph)	0.16	0.79	0.25	N/A
No Merged Embedding (A)	0.16	0.59	0.25	- 25.3%
No Edge Sampling (B)	0.16	0.65	0.25	- 17.7%
No Graph Learning (C)	0.16	0.51	0.24	- 35.4%

Table 3: Discussion of generalizability of pretrained models. *The RoBERTa and CodeBERT models in this table are fine-tuned with our simplified non-linear detection head.

Approaches	Precision	Recall	F1	F1 Dif
RoBERTa* [15]	0.69	0.73	0.72	N/A
CodeBERT* [6]	0.70	0.84	0.76	N/A
ASTRO (RoBERTa)	0.65	0.86	0.74	+ 2.8%
ASTRO (CodeBERT)	0.76	0.86	0.81	+ 6.6%

a decrease of 0.20 in recall, indicating that merged embeddings maintain critical hierarchical information of program code. In (B), due to the large quantity of edges, we increase the sampling rate from 0.1% to 10% to imitate the learning of GraphSAGE without edge sampling. This shows a decrease of 0.14 in recall, indicating possible over-smoothing in graph representation learning. In (C), we directly apply 2148 dimensional Flair word embedding from the names of node types. This leads to a reduction in recall by 0.28, demonstrating the necessity of use graph embedding to improve node representation quality.

4.4 Discussion

We further discuss the generalizability of ASTRO with respect to pretrained models by comparing performance with and without our embedding in combination with a pretrained embedding. As shown in Table 3, when input to the same non-linear detection head, concatenating our embedding to RoBERTa and CodeBERT embeddings improves F1 scores by 2.8% and 6.6%, respectively, along with much higher recall. This demonstrates that our approach improves generalizability to additional code clones previously undetected by state-of-the-art models.

5 CONCLUSION

In conclusion, we present a novel approach that leverages structural information from ASTs combined with pretrained models to build a code clone detection framework suitable for industrial codebases. Our results demonstrate an improvement on generalizability, and can be used in tandem with existing language models to improve recall and F1 scores overall. We hope this work can inspire additional work addressing generalizability issues in neural clone detection, especially with respect to improving recall for software across domains, as in real-life industrial applications.

REFERENCES

- [1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.
- [2] Brenda S Baker. 1995. On finding duplication and near-duplication in large software systems. In *Proceedings of 2nd Working Conference on Reverse Engineering*. IEEE, 86–95.
- [3] Ira D Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant'Anna, and Lorraine Bier. 1998. Clone detection using abstract syntax trees. In *Proceedings. International Conference on Software Maintenance (Cat. No. 98CB36272)*. IEEE, 368–377.
- [4] Ruian Duan, Ashish Bijlani, Meng Xu, Taesoo Kim, and Wenke Lee. 2017. Identifying open-source license violation and 1-day security risk at large scale. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*. 2169–2185.
- [5] Chunrong Fang, Zixi Liu, Yangyang Shi, Jeff Huang, and Qingkai Shi. 2020. Functional code clone detection with syntax and semantics fusion learning. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 516–527.
- [6] 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. *arXiv preprint arXiv:2002.08155* (2020).
- [7] Nils Göde and Rainer Koschke. 2011. Frequency and risks of changes to clones. In *Proceedings of the 33rd International Conference on Software Engineering*. 311–320.
- [8] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).
- [9] Lingxiao Jiang, Ghassan Mishergahi, Zhendong Su, and Stephane Gloudu. 2007. Deckard: Scalable and accurate tree-based detection of code clones. In *29th International Conference on Software Engineering (ICSE'07)*. IEEE, 96–105.
- [10] HK Jnanamurthy, Raoul Jetley, Frans Henskens, David Paul, Mark Wallis, and Sithu D Sudarsan. 2019. Analysis of industrial control system software to detect semantic clones. In *2019 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 773–779.
- [11] Toshihiro Kamiya, Shinji Kusumoto, and Katsuro Inoue. 2002. CCFinder: A multilinguistic token-based code clone detection system for large scale source code. *IEEE Transactions on Software Engineering* 28, 7 (2002), 654–670.
- [12] Cory Kapsner and Michael W Godfrey. 2003. Toward a taxonomy of clones in source code: A case study. *Evolution of large scale industrial software architectures* 16 (2003), 107–113.
- [13] Seunghak Lee and Iryoung Jeong. 2005. SDD: high performance code clone detection system for large scale source code. In *Companion to the 20th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*. 140–141.
- [14] Maggie Lei, Hao Li, Ji Li, Namrata Aundhkar, and Dae-Kyoo Kim. 2022. Deep learning application on code clone detection: A review of current knowledge. *Journal of Systems and Software* 184 (2022), 111141.
- [15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [16] Manishankar Mondal, Chanchal K Roy, Md Saidur Rahman, Ripon K Saha, Jens Krinke, and Kevin A Schneider. 2012. Comparative stability of cloned and non-cloned code: An empirical study. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. 1227–1234.
- [17] Manishankar Mondal, Chanchal K Roy, and Kevin A Schneider. 2017. Bug propagation through code cloning: An empirical study. In *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 227–237.
- [18] Hitesh Sajnani. 2016. *Large-scale code clone detection*. University of California, Irvine.
- [19] Hitesh Sajnani, Vaibhav Saini, Jeffrey Svajlenko, Chanchal K Roy, and Cristina V Lopes. 2016. Sourcererc: Scaling code clone detection to big-code. In *Proceedings of the 38th International Conference on Software Engineering*. 1157–1168.
- [20] Jeffrey Svajlenko and Chanchal K Roy. 2015. Evaluating clone detection tools with bigclonebench. In *2015 IEEE international conference on software maintenance and evolution (ICSME)*. IEEE, 131–140.
- [21] Pengcheng Wang, Jeffrey Svajlenko, Yan Zhao Wu, Yun Xu, and Chanchal K Roy. 2018. CCAAligner: a token based large-gap clone detector. In *Proceedings of the 40th International Conference on Software Engineering*. 1066–1077.
- [22] Wenhan Wang, Ge Li, Bo Ma, Xin Xia, and Zhi Jin. 2020. Detecting code clones with graph neural network and flow-augmented abstract syntax tree. In *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 261–271.
- [23] Huihui Wei and Ming Li. 2017. Supervised Deep Features for Software Functional Clone Detection by Exploiting Lexical and Syntactical Information in Source Code.. In *IJCAI*. 3034–3040.
- [24] Martin White, Michele Tufano, Christopher Vendome, and Denys Poshyvanyk. 2016. Deep learning code fragments for code clone detection. In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 87–98.
- [25] Hao Yu, Wing Lam, Long Chen, Ge Li, Tao Xie, and Qianxiang Wang. 2019. Neural detection of semantic code clones via tree-based convolution. In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*. IEEE, 70–80.
- [26] Jian Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, Kaixuan Wang, and Xudong Liu. 2019. A novel neural source code representation based on abstract syntax tree. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 783–794.
- [27] Weiwei Zhang, Shengjian Guo, Hongyu Zhang, Yulei Sui, Yinxing Xue, and Yun Xu. 2021. Challenging Machine Learning-based Clone Detectors via Semantic-preserving Code Transformations. *arXiv preprint arXiv:2111.10793* (2021).
- [28] Yue Zou, Bihuan Ban, Yinxing Xue, and Yun Xu. 2020. CCGraph: a PDG-based code clone detector with approximate graph matching. In *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 931–942.