OffSide: Learning to Identify Mistakes in Boundary Conditions

Jón Arnar Briem, Jordi Smit, Hendrig Sellik, Pavel Rapoport, Georgios Gousios, Maurício Aniche
Delft University of Technology
The Netherlands
j.a.briem,j.smit-6,h.sellik,p.rapoport@student.tudelft.nl,G.Gousios,M.F.Aniche@tudelft.nl

ABSTRACT
Mistakes in boundary conditions are the cause of many bugs in software. These mistakes happen when, e.g., developers make use of ‘<’ or ‘>’ in cases where they should have used ‘<=’ or ‘>=’. Mistakes in boundary conditions are often hard to find and manually detecting them might be very time-consuming for developers. While researchers have been proposing techniques to cope with mistakes in the boundaries for a long time, the automated detection of such bugs still remains a challenge. We conjecture that, for a tool to be able to precisely identify mistakes in boundary conditions, it should be able to capture the overall context of the source code under analysis. In this work, we propose a deep learning model that learn mistakes in boundary conditions and, later, is able to identify them in unseen code snippets. We train and test a model on over 1.5 million code snippets, with and without mistakes in different boundary conditions. Our model shows an accuracy from 55% up to 87%. The model is also able to detect 24 out of 41 real-world bugs; however, with a high false positive rate. The existing state-of-the-practice linter tools are not able to detect any of the bugs. We hope this paper can pave the road towards deep learning models that will be able to support developers in detecting mistakes in boundary conditions.

KEYWORDS
software engineering, software testing, boundary testing, machine learning for software engineering, machine learning for software testing, deep learning for software testing.

1 INTRODUCTION
Software systems commonly suffer from defects that are caused by mistakes in boundary conditions. In simple words, these mistakes happen when developers use ‘<’ or ‘>’ in cases where they should have used ‘<=’ or ‘>=’ (or vice-versa).

As an example, see the two code snippets, extracted from J Pac-Man [4], an educational Java PacMan game used to teach software testing, in Listing 1. The method withinBorders() decides whether the coordinates (i.e., x and y) are within the game map. To that aim, the method performs four comparisons: whether x and y are positive numbers, and whether x and y are within the width and the height of the map. Although each of these conditions might seem simple in isolation, when put together, the expression becomes complex and prone to mistakes. A less attentive developer might, for example, write x > 0 instead of x >= 0, which would cause a bug in the game. The same type of mistake might also happen inside loops. Off-by-one errors\(^1\) happen when a loop executes one time too many or too few.

\(^1\)https://en.wikipedia.org/wiki/Off-by-one_error

\(^2\)https://github.com/serg-delft/jpacman

\[ \text{public boolean withinBorders(int x, int y)} \] 
\[ \text{\{ return } \ x \geq 0 \&\& \ x \leq \text{getWidth()} \&\& \ y \geq 0 \&\& \ y \leq \text{getHeight()} ; \} \]

\[ \text{// Incorrect code: x > 0} \]
\[ \text{public boolean withinBorders(int x, int y)} \] 
\[ \text{\{ return } \ x > 0 \&\& \ x \leq \text{getWidth()} \&\& \ y \geq 0 \&\& \ y \leq \text{getHeight()} ; \} \]

\[ \text{// Correct code: x >= 0} \]

Figure 1: An example of a mistake in a boundary condition. Code extracted from J Pacman\(^2\).

These mistakes are particularly difficult to find in source code. After all, the result of the program is not always obviously wrong, as it is “merely off by one”. In most cases, the mistake will lead to an “out of bounds” situation, which will then result in an application crash. In worst scenarios, such bugs can even lead to security breaches, as they might not crash an application, but lead to arbitrary code execution or memory corruptions, which are potentially exploitable by adversaries [3].

Not surprisingly, many software engineering researchers have been proposing (manual) boundary testing techniques (e.g., [7, 9, 15, 19, 20]) since the early days. However, manually inspecting code for off-by-one errors is very time-consuming since determining which binary operator is actually the correct one is usually heavily context-dependent. The industry has been relying on static analysis tools, such as SpotBugs or PVS-Studio. SpotBugs promises to identify possible infinite loops, as well as array indexes, offsets, lengths, and indexes that are out of bounds. PVS-Studio also tries to identify mistakes in conditional statements and indexes that are out of bounds in array manipulation. And while they can indeed find some of them, static analysis tools are known for providing a high number of false positives [5, 10, 11].\(^3\)

In this work, we conjecture that a model based on distributed path representations of methods [2] is able to learn syntactical patterns, including code context, that can lead to off-by-one errors. To examine this, we train and test OffSide, a Code2Vec-like model on over 1.5 million code snippets, mutated to include errors in boundary conditions. Our model shows an accuracy from 55% up to 87% (depending on the type of statement where the boundary error was in). We also mine and manually examine 41 real-world bugs attributable to off-by-one errors and show that our model is able to

\(^3\)Although the papers we cite did not focus specifically on the mistakes we are discussing here, we will later show in our study that these static analysis tools are not enough in detecting mistakes in boundary conditions.
identify the locations of 24 of them. Existing state-of-practice linter tools are not able to detect any of them.

The main contributions of this paper are.

- The proposal of a deep learning model for detecting of mis-takes in boundary conditions.
- A quantitative and qualitative evaluation of our model in a training set of more than 1.5 million code snippets and a testing set of 28k code snippets.
- A research agenda that aims at improving the model.

## 2 APPROACH

Path representations of graphs are a general technique for embedding graphs into lower dimensions [17]. In the context of source code, path representations work by walking the AST and encoding the sequence of nodes encountered by the walks. Those walks usually start from random terminal nodes and end at other terminal nodes. By taking various random walks on the AST and embedding the result, models, such as Code2Vec, can learn the context in which specific token sequences are found as well as an overall representation of code, given a training label.

**OffSide** is based on the premise that the attentional layer in Code2Vec will learn the appropriate weights to enable it to differentiat-e between cases where path embeddings encode errors in boundary conditions. Intuitively, to achieve maximum discriminatory power, it should be trained with path embeddings of pairs of correct and non-correct examples of the same code. Such pairs can be easily generated by introducing artificial mutation operations that inject bugs, similarly to Pradel and Shen [18].

## 3 DATA COLLECTION

To train **OffSide**, we need both positive (containing bugs) code and negative code (bug-free). We use Code2Vec’s dataset as a starting point, and create bugs by performing specially-designed mutations, inspired by Pradel and Shen [18]. In the following, we discuss each of these steps in detail.

### 3.1 Datasets

To train and validate **OffSide**, we use, as basis, the same Java-1arge dataset collected by Alon et al. [22]. It consists of 1,000 top-starred projects from GitHub and contains about 4 million code snippets. From that dataset, we filter out (1) all methods not containing a ‘<’, ‘<=’, ‘>’, or ‘>=’ (hereafter referred to as comparators), (2) constructors, and (3) files that the parser we used, JavaParser⁴, is unable to parse. This results in a base dataset of 1,357,210 methods containing a total of 2,736,573 comparators.

For further analysis of our base dataset and the performance of **OffSide** with different sub-types of boundary mistakes, we define two classifications for each comparator. The type of operator (less, lessEquals, greater, or greaterEquals) and the type of statement it occurs in (there are 14 different types: if, for, while, return, ternary operator, method, assert, do, variable declaration, assign, expression, object creation, array initialization, and cast). The type of statement containing the comparator is defined by the class of the parent node in the AST.

### 3.2 Generating Positive and Negative Instances

For each code snippet m in our dataset, we first identify all statements $S_t$ of a given type $t$, $t \in \{ if, for, while, \ldots \}$ (see the full list in Table 1). For each type of statement $t$ that exists in code snippet $m$, we randomly select one $s \in S_t$. Let $C \in \{ >, \geq, <, \leq \}$ be all the comparators that exist in the statement $s$. We randomly select one $c \in C$, and mutate it (i.e., '$<\to\leq$ and vice versa, same goes for $'>'$ and $'\geq\to\leq$). Both the original method (bug-free) and all its mutations (buggy) are added to our final preprocessed dataset.

This process results in a 54:46 proportion of positive (containing a bug) and negative (bug-free) instances during training. Overall, our training set consists of 1,512,785 data points (bug-free and buggy methods), our validation set of 28,086 data points, and our test set of 104,958 data points. The split into training, test and validation datasets is based on the split in the original Java-large dataset collected by Alon et al. [2]. Since that split is done on a project level, this leads to the test dataset only containing methods from projects that are not included in the other two datasets and vice versa.

### 3.3 Source code representation

After a method has been extracted, it passes through the same preprocessing pipeline as used in the original Code2Vec paper [2]. The source code of each method is again turned into an AST using the modified JavaExtractor from [22]. We then select at most 200 paths between 2 unique terminals in the AST of the method. We encode these terminals into integer tokens using the dictionary used by Code2Vec [2] and hash the string representation of the paths with Java hashcode method. This means that each method

---

1. https://code2vec.org/
2. https://javaparser.org/
in Java code is turned into a set of at most 200 integer tuples of the format \((\text{terminal}_i, \text{path}, \text{terminal}_j)\) whereby \(i \neq j\) and \(\text{path}\) is an existing path in the AST between source \(\text{terminal}_i\) and target \(\text{terminal}_j\).

4 MODEL

4.1 The Neural Network Architecture

OffSide is an attention-based model based on Code2Vec, whereby the overwhelming majority of the weights are in the embedding layer of the network. The architecture of the model is shown in Figure 2. The model takes a set of at most 200 integer token tuples of the format \((\text{terminal}_i, \text{path}, \text{terminal}_j)\) as an input. It embeds these inputs into a vector with 128 parameters, whereby the terminal tokens and the path tokens each have their own embedding layer. These embeddings are concatenated into a single vector and passed through a dropout layer. These 200 vectors are then combined into a single context vector using an attention mechanism. The context vector will be used to make the final prediction. We modify the output of the Code2Vec model, by adding a Sigmoid activation with a single (binary) output unit; this enables Code2Vec to classify input code as buggy or non-buggy.

4.2 Training

We use binary cross-entropy as the loss function and Adam [14] as the optimization algorithm. The training process is halted after the accuracy on the validation set did not increase for 2 epochs and the weights with the lowest validation loss were kept.

The authors of Code2Vec speculate that their pre-trained weights could be used for transfer learning [2]. We experiment with applying transfer learning in two ways. Initially, we attempt "feature extraction" whereby the pre-trained weights of the Code2Vec model were frozen and only the final layer was replaced and made trainable. Then, we try "fine-tuning" with pre-trained weights of the Code2Vec model as the initial value of our model and allow the model to update all the values as it sees fit, expect the embeddings weights. Finally, we also train a model with "randomly initialized weights" as a baseline. We base our further results on the "fine-tuning" model, as it proved best in terms of F1-score (see results in the appendix [13]).

4.3 Reproducibility

All the source code and datasets can be found in our code repository.5 The dataset we used can be found on Zenodo.6

5 EVALUATION

We define two research questions to evaluate OffSide’s performance:

RQ1. What is the accuracy of OffSide in detecting errors in boundary conditions?

RQ2. How does OffSide perform on real-world errors in boundary conditions, when compared against the state-of-the-practice linting tools?

To answer RQ1, we apply OffSide on the test set composed of 104,958 data points (see Section 3.2). We evaluate the performance of the model by means of traditional binary classification performance metrics, such as precision, recall, accuracy, and F1.

To answer RQ2, we apply OffSide on 41 real-world bugs in boundary conditions that we extract from the 500 most starred GitHub Java projects. The analyzed projects are not part of our training and test sets before, and thus are never seen by our model. By means of a Pydriller script [21], we extract a list of candidate commits where authors made a change in comparators (e.g., a ‘>’ by a ‘<’ or ‘<=’). This process returns a list of 1,571 candidate commits. We then 1) randomly select one candidate commit, 2) manually analyze the change and the commit message to identify whether it is a bug fix, 3) apply our model to the buggy (where we expect the model to identify an error) and to the bug-free version (where we expect our model to not predict an error) of the code, and 4)

5https://github.com/SERG-Delft/ml4se-offside

6https://zenodo.org/record/3606812
Table 2: Performance metrics of our model per statement type.

<table>
<thead>
<tr>
<th>Statement type</th>
<th>Total</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>If</td>
<td>49,418</td>
<td>0.72</td>
<td>0.6666</td>
<td>0.7462</td>
<td>0.7042</td>
</tr>
<tr>
<td>For</td>
<td>39,018</td>
<td>0.8723</td>
<td>0.8565</td>
<td>0.8844</td>
<td>0.8702</td>
</tr>
<tr>
<td>While</td>
<td>5,718</td>
<td>0.7272</td>
<td>0.6691</td>
<td>0.757</td>
<td>0.7104</td>
</tr>
<tr>
<td>Return</td>
<td>3,558</td>
<td>0.7476</td>
<td>0.6931</td>
<td>0.7779</td>
<td>0.7331</td>
</tr>
<tr>
<td>Ternary</td>
<td>3,114</td>
<td>0.6574</td>
<td>0.5466</td>
<td>0.7021</td>
<td>0.6147</td>
</tr>
<tr>
<td>Method</td>
<td>1,954</td>
<td>0.6592</td>
<td>0.5916</td>
<td>0.684</td>
<td>0.6345</td>
</tr>
<tr>
<td>Assert</td>
<td>608</td>
<td>0.6135</td>
<td>0.5164</td>
<td>0.6408</td>
<td>0.5719</td>
</tr>
<tr>
<td>Do</td>
<td>598</td>
<td>0.689</td>
<td>0.5886</td>
<td>0.7364</td>
<td>0.6543</td>
</tr>
<tr>
<td>Var. decl.</td>
<td>544</td>
<td>0.6581</td>
<td>0.5368</td>
<td>0.7087</td>
<td>0.6109</td>
</tr>
<tr>
<td>Assign</td>
<td>336</td>
<td>0.5982</td>
<td>0.4464</td>
<td>0.641</td>
<td>0.5263</td>
</tr>
<tr>
<td>Expression</td>
<td>72</td>
<td>0.6111</td>
<td>0.4167</td>
<td>0.6818</td>
<td>0.5172</td>
</tr>
<tr>
<td>Obj. creation</td>
<td>20</td>
<td>0.55</td>
<td>0.4</td>
<td>0.5714</td>
<td>0.4706</td>
</tr>
<tr>
<td>Total</td>
<td>104,958</td>
<td>0.7733</td>
<td>0.7303</td>
<td>0.799</td>
<td>0.7631</td>
</tr>
</tbody>
</table>

We apply state-of-the-practice linting tools. We stop this process after identifying 41 bugs, which we deem satisfactory for this analysis.7

We use three different static analyzers as a baseline for our evaluation. SpotBugs (v.4.0.0-beta1)8, formerly known as FindBugs[8], is an open-source static code analyzer for Java. It analyzes Java bytecode for occurrences of different patterns that are likely containing a bug. At the time of writing this report, SpotBugs is able to identify over 400 of such patterns9, out of which six we consider to be relevant for the type of bug we are detecting: IL_INFINITE_LOOP (an apparent infinite loop), RpC_REPEATED_CONDITIONAL_TEST (repeated conditional tests), RANGE_ARRAY_INDEX (array index is out of bounds), RANGE_ARRAY_OFFSET (array offset is out of bounds), RANGE_ARRAY_LENGTH (array length is out of bounds), and RANGE_STRING_INDEX (string index is out of bounds). We also use PVS-Studio (v.7.04.34029)10, which is a proprietary static code analysis tool for programs written in C, C++, C# and Java. Out of 75 possible patterns for Java code analysis, two were suitable for our evaluation: V6003 (potential error in a construct consisting of conditional statements), and V6025 (index value is outbound the valid range). Finally, we use the static analyzer integrated into IntelliJ IDEA Ultimate11 (v. 2019.2.3), a popular IDE among Java developers.

5.1 RQ1: What is the accuracy of OffSide in detecting errors in boundary conditions?

In Table 2, we show the performance metrics of our models, grouped by statement type and comparator, respectively. In the appendix [13], we discriminate the performance of each statement type and its all possible comparators.

The average accuracy of the model is around 0.65 to 0.75. Overall, we observe that our models provided a median accuracy of 0.67 (when looking at the accuracy numbers per statement, Table 2) and 0.75 (when looking at the accuracy numbers per comparator, see Appendix). Moreover, the F1 measures (which, different from the accuracy metric, takes both false positives and false negatives into account) present similar numbers.

The accuracy per statement type is correlated to the amount of training data. We observe that the precision is correlated with the total amount of data points available for each context type and the types which have the highest number of occurrences also tend to produce a higher F1 score. For example, our model achieves an F1 score of 0.87 when detecting bugs in for loops, which are well represented in our dataset. However, our model only achieves an F1 score of 0.52 detecting bugs when assigning a boolean value to a variable with a logical condition, a case that is severely underrepresented in the training data.

We observe that the model can also perform well with boundary mistakes in moderately underrepresented classes such as return statements (F1 score of 0.73) and while loops (F1 score of 0.71). This might indicate that the problems were similar enough for the errors in if statements and for loops for the model to generalize. We conjecture that the most underrepresented classes like assigning value to a variable are noisy and the model was not able to generalize towards those classes.

The model is biased towards for loops. A detailed analysis shows that the classical for loop (with < operator) scores are significantly higher than others (86-89% accuracy). This might be due to for loops with comparators such as (int i = 0; i < number; i++) being considered as a boilerplate in Java code. We also observe that the model is biased towards predicting <= in a for loop as a bug and < not as a bug. This can be explained by the balance of the training set where the majority of the for loops contain a < operator. Hence, we conjecture that the model learns to classify our mutated code with <= operator as faulty.

Training on all types of statements seems to benefit specific statements. We were also surprised by the positive results in if statements, as the model achieved an accuracy and F1 score of 0.72 and 0.70, respectively. If statements have no default structure, as for loops have, and thus, knowledge of the context is needed in order to make a good prediction. To better understand our results, we trained a model specifically for if conditions to see how it would perform on the test data. 12 Interestingly, the results are not better than the model trained with all mutations and the model performs worse when judging if statements that contain a > operator (F1 score 0.63 vs 0.34). One possible reason for this might be that the model can generalize the relationship better with more data, independent of the contexts like if conditions or for loops.

**RQ1 summary:** Deep learning models can learn how to identify mistakes in boundary conditions. Our initial models provides an average accuracy of 65% to 75%. The amount (and quality) of training data plays an important role in the process.

---

7 We specifically focused on for loops in the last three bugs, as to increase its presence in our dataset.
8 SpotBugs official GitHub page: https://github.com/spotbugs/spotbugs
10 PVS-Studio official home page: https://www.viva64.com/en/pvs-studio/
11 https://www.jetbrains.com/idea/
12 The results of this specific model can be seen in our appendix [13].
5.2 RQ2: How does OffSide perform on real-world errors in boundary conditions, when compared against the state-of-the-practice linting tools?

We present the performance of OffSide in the 41 real-world boundary mistakes (both in the buggy version as well as in the fixed version of the code snippet) in Table 3.

The model is able to detect real-world bugs, but with a high false positive rate. We observe that our tool was correctly able to identify the bug in 24 out of the 41 cases, or 58% of the cases. Out of these 24 correct predictions, the tool was able to also correctly label the non-buggy version of the code as bug-free 16 times (66% of the 24 correct predictions, or 39% of the 41 total cases).

The state-of-the-practice linting tools did not find any of the real-world bugs. We also note that these bugs are indeed not trivial to be identified. In fact, they were not identified by any of the state-of-the-practice linting tools. Therefore, we did not include their results in the table.

RQ2 summary: The model presents a reasonable performance in real-world bugs. The current false positive rate, however, is high. The state-of-the-practice linters, on the other hand, do not identify such bugs.

6 DISCUSSION AND FUTURE WORK

We consider our initial exploration on whether deep learning models can learn to identify mistakes in boundary conditions to be successful. Our models had an average accuracy of 0.75 and, in real-world bugs from projects never seen before by the model, an accuracy of almost 60%. However, much still needs to be before this model can be used in the real-world:

A better, larger, and more balanced dataset. In the first version of the model, we had a training dataset of around 1.5M methods. Other similar papers make use or larger datasets, e.g., around 12M snippets in Code2Vec’s paper [2] and around 5M snippets in DeepBugs [18]. Moreover, we can see in Table 2 that elements with more data present higher accuracy. We therefore expect that a larger dataset will naturally give us better accuracy.

Moreover, as we show in Section 3.1, our initial dataset contains four times fewer usages of $\geq$ or $\leq$ compared to usages of $>$ or $<$. We also have a larger number of $i < x$, rather than $\leq$, $>$, or $\geq$, inside for loops. These differences can lead to biased training and, as a result, models tending to give false positive results in case of $\geq$ or $\leq$. One way to mitigate the issue is to create a balanced dataset with a more equal distribution of binary operators, as well as the distribution of the places of their occurrence (if-conditions, for-and while-loops, ternary expressions, etc.) In addition, while the amount of method duplication is small in Code2Vec’s dataset [1], removing duplicates might improve our accuracy; and while our approach does not currently suffer from equivalent mutants [6], this is indeed a point of attention for the future when we will make use of more advanced mutation techniques.

Finally, while we identified 41 real-world off-by-one-bugs, a larger dataset of real bugs might be of great importance in evaluating the accuracy of our model. The development (or even the reuse) of datasets like Defects4J [12] is, thus, an important future work.

Support inter-procedural analysis. Currently, we only analyze one method at a time. However, to detect more complex bugs in boundary conditions, we conjecture that this knowledge is essential, and omitting the contents of called methods might lead to unpredictable results. We plan to explore techniques that would
allow us to expand the context of the code representation, as in Li et al. [16].

Better understand the differences in RQ1 and RQ2. Interestingly, our accuracy was higher in the results of our RQ1 than in RQ2. By manually analyzing the 41 bugs we selected for RQ2, we see no reason for the model to have a lower accuracy. We intend to better understand the root cause of this difference in future work.

Fine-grained bug identification. Our model currently predicts boundary mistakes at method-level. That is, the model predicts whether a method contains a boundary mistake or not. It is not able to detect exactly which boundary is buggy. While we argue that this is already helpful for developers (developers could, for example, increase their code review efforts in methods marked as buggy by our model), future work should focus on identifying, in a fine-grained manner, where the bug is.

The generalization to other languages and problems. Mistakes in boundary conditions might happen in source code of any programming language. We wonder how much our approach would generalize to other (not only static, but also dynamic) languages and, whether deep learning models would be able to learn such bugs. Moreover, right now we solely focus on off-by-one errors. Future work can explore the effectiveness of our models in detecting any types mistakes in boundaries (e.g., not only detect a > that was mistakenly a >=, but also mistakenly a <, <=, ==, or #).

6.1 Threats to Validity

Internal validity. We consider a boundary to happen in "any existing inequality expression in the source code", i.e., any source code construct that makes use of >, >=, <, and <=. Given that our method makes use of mutations to generate buggy instances, and therefore, a clear limitation of our model is that it never learns to detect the absence of a boundary check. We expect future researchers to work on models that detect not only "bugs in boundaries that are already expressed in the source code", but also "missing boundary checks".

External validity. More replications of this work (including industry systems) are required before we claim generalizability.

7 RELATED WORK

DeepBugs by Pradel et al. [18] uses a deep learning model to identify bugs related to swapped function arguments, wrong binary operators and wrong operands in binary operation. DeepBugs feeds method and argument names into Word2Vec to calculate identifier embeddings and then aggregates those in a feed-forward neural network that predicts whether a code is buggy or not. Interestingly, the model is trained on artificially created bugs.

8 CONCLUSION

Developers are prone to make mistakes in boundary conditions. Offering them mechanisms to detect such mistakes before they become bugs in their software systems is fundamental. Given that such mistakes are highly dependent on context, traditional static analysis tools currently do not yield accurate results. This paper explores the idea of identifying such mistakes by means of learning deep models. Our initial exploration shows promising results.

REFERENCES