Context-Aware Patch Generation for Better Automated Program Repair

Ming Wen1, Junjie Chen2,3, Rongxin Wu1, Dan Hao2,3, Shing-Chi Cheung1
1Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Hong Kong, China
{mwenaa,wurongxin,scc}@cse.ust.hk
2Key Laboratory of High Confidence Software Technologies (Peking University), MoE
3Institute of Software, EECS, Peking University, Beijing, 100871, China
{chenjunjie,haodan}@pku.edu.cn

ABSTRACT
The effectiveness of search-based automated program repair is limited in the number of correct patches that can be successfully generated. There are two causes of such limitation. First, the search space does not contain the correct patch. Second, the search space is huge and therefore the correct patch cannot be generated (i.e., correct patches are either generated after incorrect plausible ones or not generated within the time budget).

To increase the likelihood of including the correct patches in the search space, we propose to work at a fine granularity in terms of AST nodes. This, however, will further enlarge the search space, increasing the challenge to find the correct patches. We address the challenge by devising a strategy to prioritize the candidate patches based on their likelihood of being correct. Specifically, we study the use of AST nodes’ context information to estimate the likelihood.

In this paper, we propose CapGen, a context-aware patch generation technique. The novelty which allows CapGen to produce more correct patches lies in three aspects: (1) The fine-granularity design enables it to find more correct fixing ingredients; (2) The context-aware prioritization of mutation operators enables it to constrain the search space; (3) Three context-aware models enable it to rank correct patches at high positions before incorrect plausible ones. We evaluate CapGen on Defects4J and compare it with the state-of-the-art program repair techniques. Our evaluation shows that CapGen outperforms and complements existing techniques. CapGen achieves a high precision of 84.00% and can prioritize the correct patches before 98.78% of the incorrect plausible ones.

CCS CONCEPTS
→ Software and its engineering → Error handling and recovery; Software testing and debugging;

KEYWORDS
Context-Aware, Automated Program Repair, Patch Prioritization
to generate patches. However, doing so will further increase the search space and hence aggravate the second limitation. This issue was also observed by Barr et al. [4] who found that the genetic programming search algorithm is always bogged down within a few generations due to the search space explosion when trying out GenProg [20] with fixing ingredients at the EXPRESSION level. Therefore, better strategies are required to prioritize the search space derived from fine fixing ingredients.

In pursuit of an efficient search strategy, we analyze the patch generation process of search-based APR and identify the major factors contributing to the search-space explosion problem [21]. An APR process typically involves generating three spaces. First, it selects a buggy location generated by FL techniques (i.e., fault space). Second, it selects a mutation operator from a predefined set of operations (i.e., operation space) that can be applied to the selected buggy location. Third, it selects a fixing ingredient (i.e., ingredient space) if it is required by the mutation operator. The ingredient space can be a set of code fragments extracted from the same application [20, 23, 36, 45] or across multiple applications [41]. Therefore, the fault space, operation space and ingredient space jointly attribute to the search space explosion problem [21]. Among the three spaces, the fault space is a set of suspicious code elements, which are potentially buggy, derived from a given set of test results. It depends largely on the adequacy of these tests [52, 56]. Following existing practices, we directly leverage the suspicious values returned by FL techniques to prioritize the fault space [15, 16, 20, 33, 36, 43, 46, 48].

In this paper, we approach the search space explosion problem by studying the other two spaces (i.e., operation space and ingredient space), which collectively referred to as the fix space [21]. Specifically, we study how to efficiently search for a correct patch, if any, in the fix space of a selected suspicious code element. Existing strategies of search-based APR randomly choose operators and ingredients from the fix space [16, 19, 20, 27, 36, 46]. However, this random strategy is inefficient and subject to high possibilities of producing plausible patches [24, 25, 53]. Therefore, we propose to prioritize the search for correct patches over plausible but incorrect ones based on mutation operators’ and fixing ingredients’ likelihood to correctly repair the selected suspicious code element.

We find that the context information of the suspicious code and the ingredients can provide rich information about their likelihood to correctly repair the bug. Our finding is supported by our empirical study showing that the correct fixing ingredients and the buggy code elements share high similarity in terms of their contexts. Leveraging the findings, we develop an approach called CapGen, which stands for Context-Aware Patch Generation. CapGen works on a fine-granularity at the AST node level. It is context-aware in two aspects. First, the mutation operators are prioritized concerning the context information. Specifically, CapGen considers the AST node types of the required ingredient and the suspicious code element when selecting mutation operators. However, whether bugs can be fixed more often under certain contexts than the others remains unknown. To address this challenge, CapGen leverages substantial real bug fixes in open source projects to guide the discovery of such contexts. Second, the fixing ingredients are extracted together with their context information. Specifically, we proposed three models to capture the context information in terms of the ingredients’ syntactics and semantics. CapGen leverages these models to guide the application of the ingredients to the correct locations. This is based on our intuition that a fixing ingredient should be applied to the locations with similar contexts compared with the location where it is extracted. To the best of our knowledge, incorporating the context information when selecting mutation operators and prioritizing fixing ingredients is new to program repair.

We evaluate CapGen on the Defects4J [14] benchmark. The evaluation results show that CapGen makes two major accomplishments. First, CapGen generates plausible patches for 25 bugs and 21 of them are correct, thus achieving a high precision of 84.00%. It outperforms the current state-of-the-art approaches in terms of the number of bugs repaired correctly and the precision. More specifically, using the context models, the ranks of correct patches has been improved for 85.31% and the number of ties of the rankings has been reduced by 97.59% on average. This significantly alleviates the search-space explosion problem. Second, it generates no incorrect plausible patches for 60.00% (15/25) of the bugs. For those with incorrect plausible patches generated, the correct patches can be ranked in prior to 98.78% of the incorrect ones. It is because that our context-aware models can precisely identify the correct fixing ingredients and thus reduce the possibility of generating incorrect plausible patches. This also significantly alleviates the overfitting problem [42]. Among these 21 bugs correctly repaired, 15 of them have never been repaired by existing APR approaches, showing that CapGen well complements existing approaches. We also evaluated CapGen on the IntroClassJava [10] benchmark, the results of which show the generality of CapGen.

In summary, our paper makes the following novel contributions:

- We propose CapGen, which is a context-aware patch generation technique. To the best of our knowledge, CapGen is the first approach to incorporating the context information to prioritize mutation operators and apply fixing ingredients.
- We design three novel models to capture the context information of fixing ingredients. We also present empirical evidences supporting that the correct fixing ingredients should share high similarities with the buggy element in terms of their contexts.
- We evaluate CapGen on the benchmark Defects4J. The results show that CapGen can generate correct patches with a high precision of 84.00%. Moreover, it outperforms and complements existing state-of-the-art approaches.

### 2 BACKGROUND AND MOTIVATION

To address the two aforementioned limitations of search-based APR, we study the real bug fixes and make some interesting observations on patch granularities, the operation space and the ingredient space.

- **Patch Granularities:** We found that one important reason which hinders existing approaches [16, 20, 29] to generate the correct patches in the search space is that the designed mutation operators work at the statement level, which are too coarse-grained. Martinez et al. reported that code redundancy is significantly higher at a finer granularity than statement level [30]. This suggests that it is more likely to find good fixing ingredients at a finer granularity. Figure 1 shows an example, which is a bug caused by a wrong expression in the return statement at line 417. To fix it, developers changed the original expression to `equal1s(x, y, 1)`. The fixing ingredient `equal1s(x, y, 1)` exists at two other places (line 422 and line 442) in the same program. If we apply mutation operators at the
statement level (e.g., replacing the return statement with another) as existing approaches [16, 19, 20, 29], we cannot fix the bug by replacing line 417 with line 422 nor line 442. Another example is shown in Figure 2, where the fixing ingredient next(pos) also exists at a granularity finer than statements. These motivate us to extract fixing ingredients and apply mutation operators at a finer granularity, such as the expression level.

```
417: return equals(x, y, 1); // Fixed statement
417: return equals(x, y, 1); // Fixed statement
```

**Figure 1:** Buggy statement, fixed statement and the fixing ingredients of bug Math 63 in Defects4J.

Applying mutation operators with ingredients at a finer granularity increases the likelihood to include the correct patches in the search space [4, 30]. However, a serious side-effect is that it will increase the fix space (i.e., operation and ingredient space) significantly, exacerbating the search space explosion problem [4]. A recommended way to address this is to prioritize the patches based on their likelihood to repair a bug correctly as adopted by recent approaches [19, 53, 55]. We make two key observations on the fix space from existing literature and the motivating examples, which shed lights on how to prioritize the search of correct patches.

**Observations on the Ingredient Space:** We observe that the context information of the fixing ingredients can guide us to select the correct ingredient to the suspicious buggy location. Consider the return expression at line 417 of the example in Figure 1. Even we have decided to patch the expression using method invocations as aforementioned, we still need to select one of the many method invocations in the program to generate a patch. A random selection is less likely to generate a correct patch. The correct fixing ingredient exists at two other places in the program. If looking at their contexts (e.g., the enclosing statements where they are extracted), we can find that both of them are used in return statements, and there are no other uses of equals(x, y, 1) in this program. Therefore, this fixing ingredient should be preferentially selected over the others to patch the buggy return statement at line 417. A similar case is found for the example in Figure 2. The ingredient next(pos) is extracted from a statement enclosed by an if statement, and thus it should be preferentially selected and inserted under the buggy if statement at line 421.

This motivates us to consider the similarity between the context where an ingredient is extracted and the context where the buggy code resides when generating patches. An ingredient with a higher context similarity should be ranked higher. In section 3.2.1, we propose three models to measure the context similarities in terms of an ingredient’s and a buggy code’s genealogical structures, accessed variables and semantic dependencies. To validate if such measurement can reliably indicate the likelihood of generating correct patches, we conduct an empirical study to collect the supporting evidence in Section 3.2.2.

3 APPROACH

We propose a Context-Aware Patch Generation (CapGen) approach, which works on the Abstract Syntax Tree (AST) of a program. Specifically, all the AST nodes in the category of Type, Expression and Statement [39] are extracted as fixing ingredients. This section first introduces the context-aware prioritization of operations and the fixing ingredients. After that, it introduces how CapGen integrates the fault space, operation space and ingredient space together to prioritize the generated patches.

3.1 Context-Aware Operators Selection

Motivated by our observations, we consider the three basic AST mutation operators, which take a fixing ingredient as a source node and the place (i.e., the buggy code element) where the ingredient is to be applied as a target node.

---

1 Type refers to the AST node type in this study, such as **Infix Expression**

Deletion: delete

Reduction of the space is a critical issue in APR [21, 24], and we that of the target node, the mutated program will not be compiled because if the datatype of a fixe ingredient does not match node, rather than its nodetype, to infer context information. It is for aceptedaugmentedmutationoperator and its frequency, respectively. We use M and Freq(M) to denote a selected augmented mutation operator and its frequency, respectively. For replacement, we utilize the data type (e.g., int, boolean) of a node, rather than its node type, to infer context information. It is because that if the data type of a fixing ingredient does not match that of the target node, the mutated program will not be compiled correctly. For example, both the data types of the fixing ingredient (i.e., equals(x, y, 1)) and the target node (i.e., the returning expression) in Figure 1 are boolean.

3.2 Context-Aware Ingredients Prioritization

Existing studies have reported that a significant portion of the fixing ingredients can be found in the same file of the buggy code [4, 30]. Therefore, we extract the ingredients from the file where the target node (i.e., the buggy code element) resides. Unlike existing studies [16, 19, 20, 29, 36, 46], we also extract the context information for each ingredient. Our hypothesis is that the correct fixing ingredients should share high similarity with the target nodes in terms of their contexts. We denote the similarity as Simi(Nt, N). In the following, we first introduce three aspects of context information embedded in an AST node and its surrounding codes, and then explain how to measure the context similarity between two nodes for each aspect. After that, we validate our hypothesis via an empirical study on the BugFixSet. Finally, we introduce how to prioritize the ingredients for the three types of augmented mutation operators.

3.2.1 Extracting Context Information.

We model the context information of an AST node from three different aspects: genealogy, variable, and dependency.

First, a node’s genealogical structure can provide useful context. For instance, the infix expression equals(x, y, 1) in Figure 1 is frequently used as the returning value in return statements. Therefore, this infix expression should have higher probability to be used as the ingredient for mutating the returning values in RETURN STATEMENT than the conditions of If STATEMENT. This information of where an ingredient is frequently used can be extracted from its genealogical structures, e.g., the Ancestor Nodes of the node.

Second, variable usages in an ingredient can provide useful context information since variables are primary components of an ingredient [55]. The example in Figure 1 shows that the fixing ingredient uses the same variables (e.g., x, y) as the target expression to be replaced. Moreover, plausible patches can be generated by simply removing functionalities via replacing complex expressions with constant values (i.e., no variable usages). Patches thus generated might pass a project’s test suite due to the weak test problem [42]. These motivate us to rank those ingredients with more similar variable usages compared with the target node at higher positions.

Third, the variables used or defined in a given node are affected by or can affect other nodes. Such semantic contexts in terms of the nodes’ dependencies are reported to be helpful in capturing the characteristics of correct patches [5, 25]. Therefore, CapGen extracts this dependency information as an aspect of a node’s context to prioritize fixing ingredients.

• Modeling Genealogy Contexts:

CapGen extracts the genealogy contexts of an AST node N via checking its Ancestor Nodes to see N is used under what types of code elements, and checking its Sibling Nodes to see N is used together with what types of code elements. For ancestor nodes, we traverse from N to its ancestors until we reach a method declaration. For Sibling Nodes, we extract those nodes of category STATEMENT and EXPRESSION within the same Block of N. We store the genealogical information in $\phi$, which counts the number of occurrences of different node types among all the ancestor and sibling nodes. Algorithm 1 shows the details of this process.
Context-Aware Patch Generation for Better Automated Program Repair

ICSE ’18, May 27-June 3, 2018, Gothenburg, Sweden

Table 1: Top replacement, insertion and deletion operators augmented with target and source nodes’ types

<table>
<thead>
<tr>
<th>Replacement</th>
<th>Source Node</th>
<th>Target Node</th>
<th>Freq</th>
<th>Replacement</th>
<th>Source Node</th>
<th>Target Node</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple_Name</td>
<td>0.0214</td>
<td>Simple_Name</td>
<td>0.0095</td>
<td>Simple_Name</td>
<td>0.0213</td>
<td>MethodInvocation</td>
<td>0.0617</td>
</tr>
<tr>
<td>Boolean_Literal</td>
<td>0.1102</td>
<td>MethodDeclaration</td>
<td>0.0013</td>
<td>MethodDeclaration</td>
<td>0.0157</td>
<td>ExpressionMethod</td>
<td>0.0914</td>
</tr>
<tr>
<td>Infix_Expression</td>
<td>0.0221</td>
<td>Boolean_Literal</td>
<td>0.0075</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>ExpressionMethod</td>
<td>0.0914</td>
</tr>
<tr>
<td>Qualified_Name</td>
<td>0.0197</td>
<td>Infix_Expression</td>
<td>0.0151</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
<tr>
<td>Conditional_Expression</td>
<td>0.0155</td>
<td>Infix_Expression</td>
<td>0.0151</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
<tr>
<td>Simple_Type</td>
<td>0.0082</td>
<td>IStatement</td>
<td>0.0141</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
<tr>
<td>Boolean_Literal</td>
<td>0.0075</td>
<td>IStatement</td>
<td>0.0141</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
<tr>
<td>Parameterized_Type</td>
<td>0.0018</td>
<td>IStatement</td>
<td>0.0141</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
<tr>
<td>Primitive_Type</td>
<td>0.0009</td>
<td>IStatement</td>
<td>0.0141</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
<tr>
<td>Assignment</td>
<td>0.0003</td>
<td>IStatement</td>
<td>0.0141</td>
<td>MethodDeclaration</td>
<td>0.0282</td>
<td>MethodInvocation</td>
<td>0.0914</td>
</tr>
</tbody>
</table>

Algorithm 1: Modeling Genealogy Context

```
input : N : A given node
output : Map φ : NodeTyple → Integer: the extracted context information
1 /* encoding ancestor nodes */
2 parent ← N
3 while parent ≠ null & parent ≠ MethodDeclaration do
4if parent ≠ Block do
5 | φ(GetNodeTyple(parent)) = 1
6 endif
7 parent = GetParentNode(parent)
8 end
9 /* encoding sibling nodes */
10 parent ← GetParentNode(N);
11 while parent ≠ Block do
12 | φ(GetNodeTyple(parent)) = 1
13 end
14 children ← parent.FilterChildrenOfType(IStatement | Expression)
15 foreach node in children do
16 | φ(GetNodeTyple(parent)) = 1
17 end
```

When applying a source node Ns with context ψs under a target node with context ψt, we compute its feasibility via comparing the similarities f(ψs, ψt) between these two contexts:

\[ f(ψs, ψt) = \frac{\sum_{t \in K} \min(ψs(t), ψt(t))}{\sum_{t \in K} ψt(t)} \]

(1)

where K denotes a set of all distinct AST node types captured by ψs. Here, Equation 1 measures the proportion of the contexts in ψs that can be covered by the contexts of the source node Ns.

**Modeling Variable Contexts:**

Given an AST node N, CAPGEN extracts a set of variables (including local variables and fields) that are accessed (i.e., read/write) in this node, which is denoted as ρ. Each element in the set ρ is denoted as e, which is a tuple (Type, Name) recording the data type and the name of the variable. We use the JACCARD distance to measure the similarity between the variable contexts of two nodes ϑT and ϑs. Besides, to avoid plausible patches generated by involving complex ingredients, we give more weights to those complex ingredients.

Therefore, we use the number of variables involved in Ns to represent the weight of the ingredient.

\[ g(ϑs, ϑT) = |ϑs| \cdot |ϑT| / |ϑs \cap ϑT| \]

(2)

Two elements are the same only if both the type and the name are matched. However, for nodes like SIMPLE_NAME (i.e., variables), when we replace it with another, their names must be different. Therefore, we only require the data type is the same for such cases.

**Modeling Dependency Contexts:**

Given an AST node N, CAPGEN extracts its dependency contexts via investigating those nodes affecting N and those nodes affected by N. Specifically, we conduct intra-procedure backward slicing and forward slicing based on the definition and use (also known as def-use) of variables extracted from a given node [47]. Lines 2 to 11 in Algorithm 2 describe the details of extracting the context information by leveraging backward slicing. For each of the variables used in the node, CAPGEN slices a set of statements which affect the values of the variable. We then extract the AST nodes in these statements that are instances of Expression or Statement and store such information in a map ψ to count the number of occurrences of each node type in the sliced statements. Such context information is also extracted by leveraging forward slicing.

Algorithm 2: Modeling Dependency Contexts

```
input : N : A given node
output : Map ψ : NodeTyple → Integer: the extracted context information
1 /* context information extracted from backward slicing */
2 variableSet ← GetUseVariable(N)
3 foreach variable ∈ variableSet do
4 contextStatements ← BackwardSlicing(variable)
5 foreach stmt ∈ contextStatements do
6 ψ(stmt) = stmt.FilterChildrenOfType(IStatement | Expression)
7 foreach node ∈ stmt do
8 | ψ(GetNodeTyple(node)) = 1
9 end
10 end
11 end
```

The feasibility of applying a source node Ns with dependency context ψs under a target node with context ψt is similar to how we measure the feasibility in terms of the genealogy contexts. Specifically, we measure the proportion of the target node’ contexts that can be covered by the contexts of the source node Ns.

\[ f(ψs, ψt) = \frac{\sum_{t \in K} \min(ψs(t), ψt(t))}{\sum_{t \in K} ψt(t)} \]

(3)

where K denotes a set of all AST node types captured by ψt.

3.2.2 Empirical Evidences

We designed three different models to capture the context similarities between the fixing ingredients Ns and the target node Nt. Our intuition is that the required fixing ingredients are supposed to share high similarities with the target node to be mutated in terms of their contexts. To validate the intuition, we collected empirical evidences from the BugFixSet dataset. Among 12.65% of the bug fixes in BugFixSet, fixing ingredients exist in their associated buggy program. This proportion is consistent with existing empirical studies [3, 30]. For each of these bug fixes, we first extract the correct fixing ingredient as Ns together with its context information from the associated buggy program. We then compare it with the target node Nt at the buggy location for each of the three aforementioned context models and denote the similarity as S. For comparison, we also randomly select another ingredient Nt whose type is the same as Ns and can also be applied to the target location. For example, if Ns is a node of InfixExpression with type boolean, we also randomly select another Infix expression with type boolean whose variables can be resolved at the target location. This process is similar to the process of selecting ingredients randomly by existing approaches [16, 20, 36, 38]. We denote the context similarity between Ns and Nt as St. For each bug, we
We integrate the three models to prioritize the extracted ingredients. For deletion, the similarity between ingredients will be affected by the context similarity between the source node and the target node. Therefore, for deletion, $Simi_{Del}(N_s, N_t)$ is defined as follows:

$$Simi_{Del}(N_s, N_t) = (1 - f(\phi_s, \phi_t)) * (1 - f(\psi_s, \psi_t))$$  \hspace{1cm} (6)

where $N_o \equiv N_s$. If there is no other nodes which are identical to $N_t$, then $Simi(N_s, N_t)$ would simply be 1.

### 3.3 Patch Prioritization

In summary, to generate candidate patches, CapGen first selects a buggy node $N_s$, prioritized by its suspicious value $FL(N_s)$ generated by FL techniques. Note that existing FL techniques assign suspicious values at the line level, CapGenX applies these FL techniques by assigning a line’s suspicious value to each of its AST nodes in the same line. The suspicious value of those AST nodes (e.g., If Statement) that span across multiple lines are averaged over these lines. Then, for the selected $N_t$, CapGen selects a set of mutation operators that can be applied to $N_t$ and each mutation operator $M$ has the corresponding operation probability $Freq(M)$ by considering the contexts of $N_t$. Afterwards, CapGen prioritizes all compatible source nodes $N_s$, which are used as ingredients required by $M$, based on their context similarity $Simi(N_s, N_t)$ compared with the target node $N_t$. Finally, patches are generated, and each of them is denoted as $P = \langle N_s, M, N_t \rangle$. CapGen prioritizes those patches by integrating the fault space, the operation space and the ingredient space together. The final score of $P$ is denoted as $Score(N_s, M, N_t)$, which is calculated as shown in Equation 7.

$$Score(N_s, M, N_t) = FL(N_s) * Freq(M) * Simi(N_s, N_t)$$  \hspace{1cm} (7)

Currently, we use multiplication to combine all the models to filter out those patches generated with ingredients sharing no context similarities. Designing different ways to combine those components together is left for future work as discussed in Section 5.1.

### 4 EVALUATION

#### 4.1 Research Questions

We implemented CapGen in Java based on a code transformation library Spoon [34] and the fault localization tool GZoltar [2]. Specifically, CapGen leverages GZoltar with the algorithm Ochiai [37] to retrieve the fault space and the suspicious values. We use the program analysis tool Understand [3] for code slicing. Our experiments are run on a CentOS server with 2x Intel Xeon E5-2450 Core CPU @2.1GHz and 192GB physical memory. We set the time budget for each bug as 90 minutes following existing practices [16, 19]. Our evaluation aims to answer the following research questions:

- **RQ1**: How effectively does CapGen fix real world bugs?
- **RQ2**: Can CapGen outperform existing approaches?
- **RQ3**: How is the contribution of each model used by CapGen?

The implementation of CapGen and the details of our evaluation results are publicly available at:


#### 4.2 The Subject Programs

To answer the three RQs, we run CapGen on the benchmark datasets for program repair: DEFECTS4J [14]. The benchmark was built to facilitate controlled experiments in software debugging and testing researches [14], which has been widely adopted by recent studies on APR [19, 27, 29, 40, 54]. Following existing practices [27, 40, 53], we use DEFECTS4J of version 1.0.1 and exclude the Closure Compiler.
Table 2: Subjects for evaluation

<table>
<thead>
<tr>
<th>Subject</th>
<th>#Bugs</th>
<th>KLOC</th>
<th>Test KLOC</th>
<th>#Test Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commons Lang</td>
<td>65</td>
<td>22</td>
<td>6</td>
<td>2,245</td>
</tr>
<tr>
<td>JFreeChart</td>
<td>26</td>
<td>96</td>
<td>50</td>
<td>2,205</td>
</tr>
<tr>
<td>Commons Math</td>
<td>106</td>
<td>85</td>
<td>19</td>
<td>3,602</td>
</tr>
<tr>
<td>Joda-Time</td>
<td>24</td>
<td>11</td>
<td>14</td>
<td>4,130</td>
</tr>
<tr>
<td>Total</td>
<td>224</td>
<td>231</td>
<td>128</td>
<td>12,182</td>
</tr>
</tbody>
</table>

project because it uses a customized testing format instead of JUnit test [53]. As a result, four projects with a total of 224 real bugs are used as subjects for experiment. Their demography is given in Table 2. Evaluations on another benchmark IntroClassJava [10, 22] are also conducted (see discussion in Section 5.2).

4.3 RQ1: Performance of CapGen on Real Bugs

To answer RQ1, we apply CapGen to the Defects4J benchmark. For each bug, we validate the generated patches prioritized by their Score([Nc, M, Nc]) one by one. We first run the failing test cases to see if they pass on the generated patch. If so, we then run the regression test cases to see whether this patch introduces new bugs. We skip to validate the next patch otherwise. CapGen keeps generating and validating the candidate patches within the time budget.

After all the patches have been validated or the budget timeout, we collect the plausible patches which pass both the failing test cases and the regression test cases. We then manually analyze these patches to see if they are semantically equivalent to the patches submitted by developers. We mark them as correct patches if they are, and incorrect plausible patches otherwise.

CapGen generates plausible patches for 25 bugs in total among all the 224 bugs. We identify the correct patches for 22 of these 25 bugs successfully. Table 3 shows the details of these 22 bugs. Column Tot shows the number of the patches generated while Crt shows the number of the correct ones among them. The total number of patches generated is affected by the size of the fault space and the size of buggy source file (since more fixing ingredients will be extracted if the buggy source file is larger and contains more statements). There are cases that CapGen generates more than one correct patch. For example, CapGen generates 2 correct patches for Math 53, which requires inserting a statement to fix the bug. We find that both the two patches are equivalent to the developers’ patch since the required statement can be inserted at two adjacent places. Columns Rank and Tie in Table 3 show the rank of the first correct patch and the number of patches, including wrong patches (i.e., those did not pass the test suite) and plausible patches, that are ranked in tie with the first correct patch respectively. The value of Rank is the rank of the first position if there are ties. Column Time shows the time required for CapGen to generate the correct patches ranges from 0.18 to 50.32 minutes. Columns P-B, P-T and P-A show the number of incorrect plausible patches grouped by their relative ranks (i.e., before, in tie with and after) compared with the correct patches. We can see that CapGen generates incorrect plausible patches for 40.00% (10/25) of the bugs. However, even though CapGen might generate incorrect plausible patches, it ranks them mostly after the correct ones. Table 3 shows that 21 out of the 22 bugs, CapGen ranks no incorrect plausible patches before or in tie with the first correct patches. The only exception is Math 80, for which CapGen generates two incorrect plausible patches that are ranked before the first correct patch.

For the 21 correctly repaired bugs, 90.48% of them require a fixing granularity finer than the statement level. Examples are given in Figures 1 and 5. Some of the cases requiring fixing ingredients with finer granularities can be handled by specifically designed mutation operators. For example, for bug Chart 1, by replacing the infix expression dataset!=null with another dataset=null, CapGen can repair it. HDRepair also repaired it by including a mutation operator that negates boolean expressions [19]. Currently, CapGen includes 30 augmented mutation operators, and seven of them can repair at least one bug correctly.

These results show the effectiveness of CapGen. Specifically, it repairs 21 bugs successfully (i.e., the first patch passing all tests is correct) and achieves a high precision of 84.00% (21/25), which outperforms the existing state-of-the-art approach ACS [53] with a precision of 78.30%. Precision is vital for APR techniques because much debugging effort would be wasted if the provided patches are plausible but incorrect [53]. This high precision is achieved via patch prioritization guided by our context-aware models. Specifically, 98.78% (164/166) of the incorrect plausible patches are ranked after the correct ones (see Section 4.5 for detailed analysis).

4.4 RQ2: Comparison with Existing Approaches

We compare CapGen with five APR approaches, ACS [53], HDRepair [19], jGenProg [27], Nopol [54] and PAR [16, 19], which have been evaluated on the Defects4J benchmark within our knowledge. Among them, ACS [53] and HDRepair [19] are the state-of-the-art approaches. Similar to CapGen, HDRepair [19] and PAR [16] set the time budget as 90 minutes. Nopol [54] and jGenProg [27] set the searching time as 3 hours. ACS [53] sets the budget as 30 minutes since it only targets at condition synthesis and thus the corresponding search space is much smaller. Table 4 shows the comparison results. The baselines’ results are directly extracted from existing literature [16, 20, 53, 54]. Compared with these techniques, CapGen outperforms all of them in terms of the number...
of correctly repaired bugs and precision. By analyzing the overlap of repaired bugs between CapGen and the compared techniques, CapGen actually complements the state-of-art techniques. 15 of the 21 bugs repaired by CapGen have never been repaired by existing approaches [16, 19, 20, 53, 54].

Six of the 21 bugs repaired by CapGen can be repaired by the state-of-the-art search-based approach HDRepair [19]. This happens when the required fixing ingredient is at the statement level (e.g., Math 53) or can be handled by specific designed mutation operations (e.g., Chart 1). The dominant reason why CapGen can repair 15 new bugs is that CapGen works at a finer granularity (e.g., Figure 2). For the 4 bugs that can be fixed by HDRepair [19] but not CapGen, the major reason is that the fixing ingredients do not exist in our searching scope. For example, for Math 34, the correct fix is to replace the method invoking expression chromosomes in chromosomes.iterator() with getChromosomes(). However, the fixing ingredient getChromosomes() does not exist in the buggy program, and thus CapGen failed to repair the bug. To handle this case, HDRepair [19] designed a mutation operator that replaces the name of a method call (or a method invoking expression) by another method name (or expression) with compatible types. By including more mutation operators or increasing the scope to extract fixing ingredients, CapGen can handle more cases (see Section 5.1 for more discussions).

Table 4: Comparisons with existing tools on DEFECTS4J

<table>
<thead>
<tr>
<th>Project</th>
<th>CapGen</th>
<th>HDRepair</th>
<th>jGenProg</th>
<th>PAR</th>
<th>ACS</th>
<th>Nopol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart</td>
<td>4</td>
<td>2/2</td>
<td>0/0</td>
<td>0/-</td>
<td>2/0</td>
<td>1/0</td>
</tr>
<tr>
<td>Lang</td>
<td>5</td>
<td>3/2</td>
<td>0/0</td>
<td>1/-</td>
<td>3/0</td>
<td>3/0</td>
</tr>
<tr>
<td>Math</td>
<td>12</td>
<td>4/2</td>
<td>5/3</td>
<td>2/-</td>
<td>12/2</td>
<td>1/1</td>
</tr>
<tr>
<td>Time</td>
<td>0</td>
<td>1/0</td>
<td>0/0</td>
<td>0/-</td>
<td>1/0</td>
<td>0/0</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>10/6</td>
<td>5/3</td>
<td>3/-</td>
<td>18/2</td>
<td>5/1</td>
</tr>
<tr>
<td>Precision</td>
<td>84.0%</td>
<td>56.5%</td>
<td>18.5%</td>
<td>-</td>
<td>78.3%</td>
<td>14.3%</td>
</tr>
</tbody>
</table>

X/Y: X is the number of bugs repaired by the approach, Y is the number of bugs that repaired by CapGen and also by the approach. ‘-’ means the results are not available due to the PAR did not present the ID of the repaired bugs.

The same case for Chart 1. However, CapGen requires the existence of the fixing ingredients in the program. Those cases whose fixing ingredients do not exist in the program might be handled by ACS.

4.5 RQ3: Contributions of Each Model

We use the average results of the 22 bugs shown in Table 3 to answer this question. Specifically, we conducted nine experiments to examine the contribution of each component of CapGen. Figure 6 shows the results. Each bar has two heights. The white part indicates the first correct patch’s average rank, and their height difference (gray colored) gives the number of other patches that are ranked in tie with the correct one. In the first setup, we use only the suspicious value of the fault space (i.e., FL(N)) to rank all the patches generated. All mutation operators, which are denoted as MO, bear the same weights, and no context models are considered. In the second setup, we study the prioritization performance with the mutation operators’ weights combined (i.e., FL(N) + Freq(M)). In the remaining seven setups, we evaluate the contribution of each context model and their combinations one by one. Specifically, we integrate a single context model to the final ranking (i.e., GC for Genealogy Context, DC for Dependency Context and VC for Variable Context) and the arbitrary combinations of them. In Figure 6, bars from position 3 to 9 show the results by combining the context models with FL and MO (e.g., VC means CapGen leverages FL, MO and the VC model to produce the final ranking).

As shown in Figure 6, each individual model contributes to improve the rank of correct patches. If we use only the FL model, the patches are simply ranked by the suspicious values of the buggy locations return by GZoltar. Due to inadequacy of test cases [42], the ranking of the correct buggy location is low and contains many ties. Besides, many possible patches can be generated for each location. These two reasons account for the poor ranks (e.g., the low rank and many ties) of the correct patches. By combining the weights of the mutation operators, CapGen can significantly break the ties and reduce the number of ties by 67.23%. The performance of the first rank has also been improved by 4.57%. By further introducing all the context models of fixing ingredients, the rank is further improved by 84.31% and the number of ties are further reduced by 97.59% with respect to the ranking of using FL(N) + Freq(M).

Let us illustrate this using the bug shown in Figure 5. The buggy expression is in an Assignment under an If Statement (line 1135). The correct fixing ingredient also exists in an Assignment and a condition expression in an If Statement. By capturing such context information, we can prioritize the correct fixing ingredients at high positions. Specifically, the correct patch is ranked at 6758 with 3767 ties if only the FL model is used. By including the MO model, it can be ranked at 5807 with 1002 ties. By including all the three
context models, it can be ranked at 221 with 35 ties, which can be searched within 30 minutes. For the rankings of each individual model and the combinations of them, the improvement is computed against the ranking of using \( FL(N_{f}) \times Freq(M) \). The results show that the three models collectively achieve the best performance, followed by the combination of the two models of DC and VC.

Another crucial function of our context-aware models is that they can prioritize the correct patches before the incorrect plausible ones. The majority of the patches generated which pass all test cases are plausible but incorrect due to the weak tests provided [42, 52, 56]. Previous techniques [16, 19, 20, 29, 36, 46] use random techniques to search among these generated patches, and therefore, they are more likely to produce the incorrect plausible patches [38]. This problem can be significantly alleviated by prioritizing the patches using our context models. Figure 7 shows the distributions of the total number of incorrect plausible patches ranked before, in tie with and after the correct patches under the nine analyses. The use of FL model alone in the first analysis leads to 80.69% of the incorrect plausible patches generated being ranked in tie with or before the first correct one. The number has been reduced by 17.18% by introducing the MO model, and 98.52% in a further step by combining our context models. Finally, only 2 plausible patches are ranked before or in tie with the correct ones. We also present the results in detail for each of our context model. For example, by integrating the GC, DC and VC model alone can reduce the number by 82.22%, 80.74% and 34.07% respectively comparing with the results generated by leveraging FL and MO.

Let us further illustrate the effectiveness of our context models by case analysis. For our motivating example shown in Figure 1, CapGen can rank the correct patch at position 4 before all other 8 plausible patches. Replacing the variable \( x \) in the function \( isNaN(x) \) with field Math\_Utils\_Two\_PI is one example of incorrect plausible patches. However, due to the low context similarity shared by \( x \) and Math\_Utils\_Two\_PI, the final probability of this patch is only 4.97\(-e4\), and thus ranked after the correct patch. Replacing \( \text{Double}\.isNaN(x) \&\& \text{Double}\.isNaN(y) \) with expression \( x==0 \) can also pass the test suite. However, our variable context model favors those ingredients which involve more similar variables with the target node. Therefore, this patch is also ranked after the correct one. Similar cases for the other bugs.

CapGen fails to rank the correct patch before all incorrect plausible ones for Math 80 shown in Figure 5. Due to weak test cases provided, 81 incorrect plausible patches have been generated and 78 of them are ranked after the correct one using our context models. However, there are still 2 exceptions. For example, replacing the variable \( n \) at line 1135 with another variable \( pingPong \) can pass all the test cases, and CapGen fails to rank this one after the correct patch. It is because these two variables are used together a lot in the program and thus sharing a high context similarity. Besides, our variable context model can not filter out the rank of this patch since it only requires their types to be the same if both the ingredient and the target node are \texttt{SIMPLE NAME} as described in Section 3.2.1. These results show that each of our proposed model can help rank the correct patches at top positions and, more importantly, before the incorrect plausible ones.

5 DISCUSSIONS

5.1 On the Extensions of CapGen

CapGen performs a single mutation to generate patches, which indicates CapGen can only target at fixing those bugs requiring a single repair action currently. However, this type of bugs is common in practice [28, 58]. For example, a existing study found that around 30% of the bugs can be fixed by a single repair action [58]. In the Defects4J benchmark, 31.70% of the bugs (71/224) can be repaired by a single repair action. Note that, patches generated via a simple repair action might also be complex. For example, CapGen inserts a whole If Statement, which across three lines, to fix bug Math 53. In the future, we will include performing mutations at multiple places in CapGen. At the same time, we will also try to reduce the increase of the search space that it incurs. One possible direction is to perform the same operation at multiple buggy target nodes simultaneously in the fault space if they are identical to each other. For example, an buggy expression \texttt{return isZero?NaN : INF} exists at multiple places for Math 46, repairing this bug requires changing all these buggy expressions to the correct ones.

In the future, we also plan to include more mutation operators to increase the chances of repairing more bugs, which could be easily integrated by design. Furthermore, CapGen currently searches only in the same program file where the target node resides to find fixing ingredients. Among the 71 bugs which require only one repair action, only 27 of them whose all required fixing ingredients can be found in the same program. Therefore, extending the searching scope is a promising extension for CapGen, which can increase the chances to find the correct fixing ingredients for more bugs [30]. For example, if we search among the whole application, we can find the ingredients for two more bugs among the 71 bugs. However, including more mutation operators and enlarging the scope to search the ingredients will inevitably increase the search space and thus make it more challenge to identify the correct patch. Therefore, better searching strategies are desired. Currently, CapGen combines FL, MO and the three context models by multiplication to prioritize the search space. In the future, CapGen will try better strategies such as learning-to-rank techniques to integrate these models.

5.2 Threats to Validity

One potential threat to validity is the generality of the benchmark Defects4J, which means that our evaluation results may not be generalized to other dataset. We planed to use BugFixSet for cross validation. However, it only provides the patch for each bug without the corresponding test cases. Therefore, it makes it impossible to validate the correctness of the patches generated by CapGen automatically. The other benchmark in Java besides Defects4J is IntroClass\_Java [10, 22] within our knowledge. To alleviate the threat of generality, we also evaluated CapGen on this benchmark. IntroClass\_Java is the Java version of the IntroClass benchmark.

![Figure 7: Total number of plausible patches](image-url)
originally proposed by Le Goues et al. [22]. This benchmark contains 297 bugs from student-written homework assignments. These bugs feature complicated fixes although they are small programs [55]. We follow exiting techniques [18, 57] to check whether a generated patch is correct. We run CapGen on this benchmark and found that it can successfully repair 25 bugs in total. The state-of-the-art semantics-based APR techniques, JFix [18] and Angelix [32], which work on the IntroClassJava benchmark, are able to repair 19 and 7 bugs respectively [18]. Specifically, CapGen can repair 14 new bugs. It is because neither Angelix nor JFix can handle floating point or strings related bugs [55] limited by the capability of the constraint solving techniques. Therefore, they have only evaluated on a subset of the IntroClassJava benchmark which only includes integer and boolean-related fixes. CapGen has no such limitations and are evaluated on the whole dataset. This result also shows that CapGen complements well to existing approaches.

Besides, CapGen is designed independent of the dataset used for evaluation. Specifically, the mutation operators are prioritized based on substantial real bug fixes extracted from open source projects. For the three context models proposed to prioritize the ingredients, we conduct an empirical study and the results shows that the correct fixing ingredients indeed share high context similarity with the target nodes measured by our models. This finding is general, and can facilitate the design of selecting fixing ingredients for both search-based and semantics-based APR techniques in the future.

6 RELATED WORK

Automated program repair techniques can be broadly classified into two categories. The first category is search-based APR [16, 19, 20, 36, 38, 43, 45, 46]. The most representative of this category is GenProg [20, 46], which searches for correct patches via genetic programming algorithm. RSRepair [36] adopts the same fault space and operation space as GenProg. It differs from GenProg in that it searches among all the candidates randomly instead of using genetic programming. Reducing the number of candidate patches and reducing the test cases required for validating a patch are two important aspects to boost the efficiency of search-based APR techniques [11]. Motivated by this, AE [45] is proposed to reduce the total cost required to find a correct patch. In order to generate candidates that are more likely to be the correct patch, PAR [16] proposes to leverage fixing templates learned from human patches to generate possible candidates. Each template is able to fix a common type of bugs. The design of mutation operators of CapGen is also guided by substantial real bug fixes. However, CapGen only leverages the syntactic information to design the mutation operators (i.e., inserting a type of code element under another type of code element). PAR designs mutation operators considering the semantic information such as adding a null pointer checker. Designing with such specific information makes PAR less generalizable than CapGen. The most recent work is HDRepair [19], which leverages historical data to search for correct patches.

The second category is semantics-based APR, which synthesizes a repair patch directly using semantic information via symbolic execution and constraint solving [9, 13, 15, 31–33, 44, 54]. Staged Program Repair (SPR) [23] is a hybrid of search-based and semantics-based automated program repair technique. It leverages traditional mutation operators to generate candidate patches [8], and it is also capable of synthesizing conditions via symbolic execution. Prophet [25] was proposed based on SPR, which is capable of prioritizing candidate patches via learning from correct patches automatically using machine learning techniques. SPR and Prophet only synthesize program elements involving conditions. Nopol is also capable of synthesizing conditions [54]. SemFix [33] is capable of synthesizing right hand side of assignments besides conditions. Angelix [32] was proposed to address the scalability issue concerning semantics-based techniques using a novel lightweight repair constraint angelic forest. S3 was recently proposed to synthesize patches leveraging program-by-examples techniques [55].

Both search-based APR and semantics-based APR techniques have their advantages and disadvantages. Search-based APR is simple, intuitive and generalizable to fix all types of bugs, which make them more effective. However, the efficiency is greatly compromised by the search space explosion problem [24]. On the other hand, semantics-based APR is more effective since the search space is more tractable by using program synthesis with restricted components. However, the effectiveness could be limited by the capability of constraint solving and program synthesis. Besides, both of these two categories suffer from the overfitting problem [42, 52].

In order to generate high-quality patches, many approaches propose to rank the patches based on their likelihood to be correct recently [7, 19, 25, 31, 53, 55]. HDRepair prioritizes patches based on their similarities with previous fix patterns that are mined from software histories [19]. Prophet [25] also compares the generated patch with existing human patches to investigate its probability of being correct but leverages different features as HDRepair [19]. ACS leverages the information mined from other projects and Java documentation to prioritize the variables and predicates used in the synthesized conditions [53]. S3 prioritizes the generated patches by comparing their similarities with the original buggy program in terms of a set of features [55]. Different from them, CapGen directly leverages the context information extracted from the buggy code elements and the fixing ingredients to prioritize the generated patches. To the best of our knowledge, incorporating the context information of fixing ingredients is new to program repair.

7 CONCLUSION

We present a novel search-based APR technique, CapGen, which works on the AST node level to increase the chances of including the correct patches in the search space. The key novelty of CapGen which allows it to generate and search the correct patches efficiently is that the patch generation process is context-aware, specifically, in the following two aspects: 1) The mutation operators are prioritized considering the context information guided by large historical data. 2) Fixing ingredients are extracted together with their context information, and such information is leveraged to prioritize the generated patches. We evaluate CapGen on DEFECTS4J, and the results show that CapGen outperforms and complements existing state-of-the-art techniques. More importantly, CapGen can achieve a precision of 84.00% and can prioritize the correct patches in prior to 98.78% of the plausible but incorrect ones.

ACKNOWLEDGMENTS

The work was supported by the Hong Kong RGC/GRF Grant No. 16209217, the MSRA Collaborative Research Award, and the National Natural Science Foundation of China under Grant No. 61522201.
Fan Long and Martin Rinard. 2016. Automatic patch generation by learning


Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics (1947), 50–60.


