Enhancements to Gramfuzz, a Grammar-based Fuzzing Tool

Su Jia Yin
UW ID: sjyin
sjyin@uwaterloo.ca

Alyssa Yiqin Huang
UW ID: y672huan
y672huan@uwaterloo.ca

Dawson(Xiaohui) Men
UW ID: xmen
xmen@uwaterloo.ca

August 20, 2020

Abstract

We enhanced `gramfuzz`, a grammar-based fuzzing tool, by designed a runner program to select the program-under-test and its grammar, feed the inputs automatically, save suspicious inputs and display the testing status in an interface. Compared to the mutation-based fuzzing tool, our enhanced grammar-based fuzzing tool penetrates deeper into benchmarks' parsers.

1 Introduction

It is difficult to explore unexpected or abnormal behaviour in software engineering, especially for closed source software.[**?**] Due to the complexity of software, it is almost unavoidable that several errors or exceptions, which may lead to economic loss or even security and safety problems, occurs in a piece of software. As Paul Strassman commented, software easily rates as the most poorly constructed, unreliable, and least maintainable technological artifacts invented by man. Thus, it is important to test a piece of software before it is used as a system component. It is also interesting to find out what kind of inputs could trigger the software-under-test to crash.

When we can not access the source codes or I/O behaviours of the target software, it is impossible to use gray-boxed or white-boxed methods, including static analysis or concolic testing. Thus, though with several disadvantages, black-box testing methods, such as fuzzing, are necessary. Among all kinds of black-box fuzzing methods, the grammar-based fuzzing method is very suitable for the scenario of using closed source software as system components. As known to be an effective technique for checking security vulnerabilities in programs, grammar-based fuzzer generates well-formatted inputs by specifying inputs via the grammar of the program-under-test. [**? ?**]. Furthermore, the application programming interface and the input format generated by other programs are already known, which limits the formats and range of inputs and decreases the cost of writing grammars.

As we noticed, most of the grammar-based fuzzing tools require the users to input the generated files to the program-under-test manually, which would be extremely inefficient when there are thousands of test cases. Thus, we designed a runner program to feed the inputs generated by fuzzer into the program-under-test automatically and display the current status with an interface.

2 Description of the method

2.1 Framework or tool chain used

2.1.1 Fuzzing tools

We choose `gramfuzz`, a open-source grammar-based fuzzing Python library, to write the grammars for benchmarks and designed an input generator. As comparision, we choose `pyZZUF`, a Python library implements bit-flip zzuf mutator, to generate inputs according to a random inputs generated by `gramfuzz`.

2.1.2 Runner program

To increase the efficiency of fuzzing test, we designed a runner program. As Fig. 1 shows, the interface of the runner program allows the user to choose the fuzzer, benchmark, input type (feed), timeout. Also, the user can select how many input files will be generated and feed to the benchamrk in a single round and how many rounds will be executed.

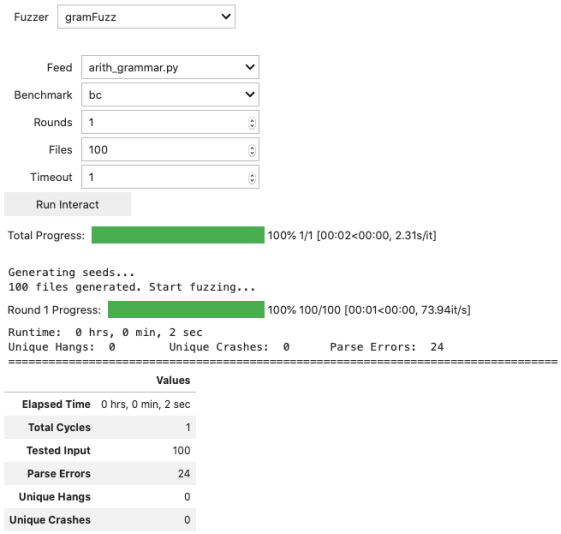


Figure 1: Interface of runner program

2.2 Benchmarks used

We chose two calculators as benchmarks. One is `bc`, another one is `calc`. Both two programs are command-line programs, and their input formats are also similar to each other. Therefore, it would be clear to illustrate the strength of grammar-based fuzzing that it can decrease the possibility of the inputs rejected by the parser and penetrate more in-depth than other fuzzing methods by generating well-formatted inputs.

Furthermore, `bc` and `calc` could be considered as representative of program-under-test, because their inputs can be divided into two types, one type is the arithmetical calculation, and the other type is program statement, including defining or calling functions, if-else conditional branchings, for loops and while loops. Thus, these two benchmarks are very typical, which makes the results have general meanings.

2.3 Grammar designed

We noticed that the average execution time for two benchmarks is significantly longer when the input is arithmetical calculation. It is because that big numbers and complicated calculations, such as exponent arithmetic, will trigger long-running uninterruptible computation. Since we detected the program hang by setting a timeout, we designed grammar for arithmetical calculation inputs and program statement inputs separately. Thus, we could set a longer timeout for arithmetical calculation when running the test. As for mutation-based fuzzer, we used seed inputs generated by the grammar-based fuzzer according to specific grammars to mutate and generate inputs for the specific type of inputs.

3 Results

3.1 Experimental setup

We run the testing on Mac OS X. Table 3.1 shows the version of software or library we used.

Table 1: Experimental setup

Software/Library	Version
`gramfuzz`	1.14.0
`pyZZUF`	0.1
`bc`	1.06
`calc`	2.12.7.1
Python	3.7.6

3.2 Experimental result

Table 3.2 shows the experimental results. We found crashes when using `gramfuzz` to test `bc` with program statement inputs. The crashes were triggered by segment fault.

Table 2: Experimental result for 100000 tested inputs

Fuzzer	Benchmark	Input type	Parser errors	Hang	Crash	Duration
`bc`	Arithmetic	`gramfuzz`	19045	0	0	00:30:52
		`pyZZUF`	50000	50000	0	14:10:24
	Statement	`gramfuzz`	9449	0	24	00:27:17
		`pyZZUF`	100000	0	0	00:51:06
`calc`	Arithmetic	`gramfuzz`	0	0	0	00:35:59
		`pyZZUF`	50000	50000	0	14:23:54
	Statement	`gramfuzz`	99745	0	0	00:47:46
		`pyZZUF`	100000	0	0	00:52:36

Compared `gramfuzz` to `pyZZUF` according to the data in Table 3.2, the results are presented as Fig. 3.2 and Fig. 3.2. From Fig. 3.2, we can concluded that it took more time to test 100 thousands input files at any cases, especially when the input tpye is arithmetic calculation. From Fig. 3.2, we could see that the inputs generated by `gramfuzz` were significantly less likely to be rejected by the parser than inputs generated by `pyZZUF`, regardless of the benchamrk and the input type.

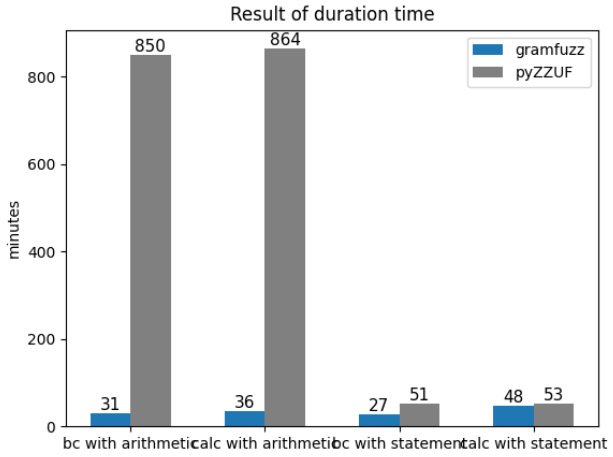
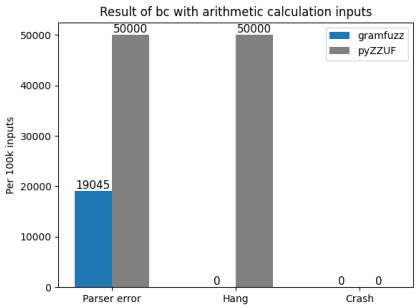
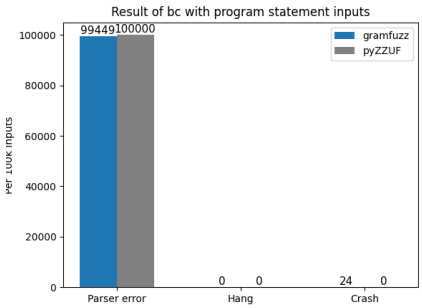


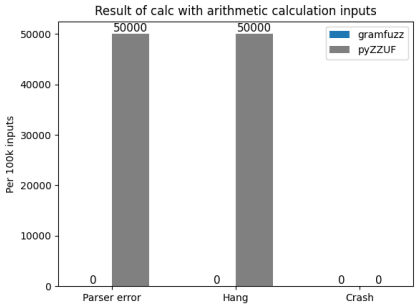
Figure 2: Duration time for testing 100000 inputs file



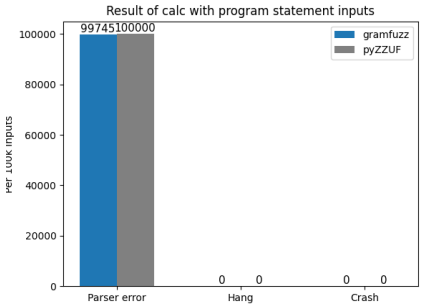
(a) `bc` with arithmetic calculation inputs



(b) `bc` with program statement inputs



(c) `calc` with arithmetic calculation inputs



(d) `calc` with program statement inputs

Figure 3: Results compared `gramfuzz` to `pyZZUF`

4 Conclusions and future work

4.1 Conclusions

4.2 Future work