SCIS 2022 2022 Symposium on Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics, Information and Communication Engineers

Efficient Machine Learning Method for Protocol Fuzzing: Improvement of Sequence-to-Sequence Model and Refined Training Data

Bo Wang * Toshihiro Maruyama * Ako Suzuki * Yuichi Kaji †

Keywords: Fuzz Testing, Protocol, Machine Learning, Seq2Seq

1 Introduction

Fuzz testing is one of software testing methods for finding software vulnerabilities and is used as a technology for finding unknown security vulnerabilities as a black-box test. Although many fuzz testing methods that are based on machine learning have been investigated, they cannot analyze and learn the real-time status of communication protocol [1]. The authors focus on efficient machine learning for protocol fuzzing. We present major problems of current fuzzing tools (fuzzers) that have not yet been solved, and introduce techniques to get around the problems.

2 Improvement of Sequence-to-Sequence model

In the conventional approach of fuzzing, we prepare massive quantity of machine-generated fuzz data, feed the data to a target software and try to detect vulnerabilities. This approach is however not suitable for protocol fuzzing because the data used in communication protocol is strongly context-sensitive. Randomly generated fuzz data often bring early termination of the communication, which makes it difficult to increase the test coverage of the target software.

For efficient protocol fuzzing, it is essential to filter out fuzz data that seems not like protocol data, and machine learning techniques such as Sequence-to-Sequence (Seq2Seq) might be contributing for the sake; we construct a model of a communication protocol by machine learning, and use the model to qualify randomly-generated fuzz data.

Seq2Seq is a commonly used technique for natural language translation [2]. The natural language translation of the Seq2Seq model is a learning model that converts an input sequence into an output sequence. In this context, the sequence is a list of symbols, corresponding to the words in a sentence. We will use the model to gain efficient protocol fuzzing, and give why Seq2Seq has been chosen.

Generally, for modeling, the natural language translation of the Seq2Seq model is processed by pairing massive input language sentences and target (output) language sentences, where sentences are encoded as a fixed-length string. However, this machine learning model is not suitable to learn communication protocols because communication data in a protocol are dynamic, long and variable both in length and in contents.

To fit Seq2Seq model to communication protocols, we consider to focus a certain part of communication data instead of trying to learn entire communication data. In the case of Bluetooth protocol, for example, the "operations field data" can be a good candidate to be learned.

Another concern of the approach is that it is not sufficient to let the model distinguish seemingly correct /incorrect protocol data. To boost the efficiency of fuzzing, we attach training data with a heuristically-determined tag that indicates if the data is likely to bring software failure or not. This makes the learned model more informative, and should contribute to improve the efficiency of protocol fuzzing.

3 Case Study and Summary

The authors applied the proposed techniques for Bluetooth protocol, where the heuristic tag for the training data are obtained from the authors' previous study. The result suggests that a model of the communication data can be constructed more efficiently than naively utilizing Seq2Seq model. The detailed method and results will be explained in the full-version of the manuscript.

References

[1] G. Saavedra, et al., "A Review of Machine Learning Applications in Fuzzing," *CoRR*, vol. abs/1906.11133, 2019.

[2] I. Sutskever, et al., "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, Curran Associates, Inc., pp. 3104–3112, 2014.

* JVCKENWOOD Corporation, 3-12, Moriyacho, Kanagawa-ku, Yokohama-shi, Kanagawa, 221-0022 Japan. (wang.bo, maruyama.toshihiro, suzuki.ako@jvckenwood.com)
† Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan. (kaji@icts.nagoya-u.ac.jp)