Programming Language Representation with Semantic-level Structure

Anonymous Author(s)

ABSTRACT

NLP (NLP) technique becomes one of the core techniques for developing text analytics applications. For developing the NLP applications, the applications are required to achieve high reliability before it goes to market. The trustworthiness of the prevalent NLP applications is obtained by measuring the accuracy of the applications on held-out dataset. However, evaluating NLP on testset does with held-out accuracy is limited to show its quality because the held-out datasets are often not comprehensive. While the behavioral testing over multiple general linguistic capabilities are employed, it relies on manually created test cases, and is still limited to measure its comprehensive performance for each linguistic capability. In this work, we introduce Fuzz-CHECKLIST, an NLP model testing methodology. Given a linguistic capability, The Fuzz-CHECKLIST finds relevant testcases to test the linguistic capability from existing datasets as seed inputs, generates sufficient number of new test cases by fuzzing the seed inputs based on their context-free grammar (CFG). We illustrate the usefulness of the Fuzz-CHECKLIST by showing input diversity and identifying critical failures in state-of-the-art models for NLP task. In our experiment, we show that the Fuzz-CHECKLIST generates more test cases with higher diversity, and finds more bugs.

ACM Reference Format:
Anonymous Author(s). 2022. Programming Language Representation with Semantic-level Structure. In *Proceedings of ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2022).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Software testing is the cruicial process when developing software. It evaluates an attribute or capability of the software and determines that it meets the requirements by examining the behavior of the software under test. Software testing in the early stage of the development finds bugs, and fixing them saves amount of costs. In addition, reliable software testing methodology ensures software quality to users in that the software meets requirements by verification and validation. Regarding that, NLP application is a branch of artifical intelligence software, and testing NLP application also becomes important process as well.

The prevalent models of NLP are evaluated via train-validation-test splits. train and validation set is used to train the NLP model and the hold-out set is used for testing by measuring accuracy. The accuracy is a indicator of the performance of the models.

Despite its usefulness, the hold-out set often overestimates the performaces. Each dataset comes with specific biases, and the biases increase the discrepancy of distribution between dataset and real-world [1]. The aforementioned accuracy on hold-out set does not consider the discrepancy and it is limited to achieve comprehensive performrance of the NLP model. As a consequence, it is difficult to analyze where the errors comes from [2].

2 BACKGROUND

3 RELATED WORK

4 APPROACH

5 RESEARCH QUESTIONS FOR EVALUATION

6 EXPERIMENT

7 RESULT

REFERENCES

[1] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5389–5400. https://proceedings.mlr.press/v97/recht19a.html

[2] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, Reproducible, and Testable Error Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 747–763. https://doi.org/10.18653/v1/P19-1073

ISSTA 2022, 18-22 July, 2022, Daejeon, South Korea

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn