PS1-23 Thu. August 17, 2023, 16:00~17:00 / Room_ B Bldg. Corridor

Proposal of Efficiency Metric for White-Box Deep Learning Testing

Joonwoo Lee [1], Hansae Ju [1] and Scott Uk-Jin Lee [2*]
[1] Major in Bio Artificial Intelligence, Dept. of Applied Artificial Intelligence, Hanyang University, Ansan, Korea
[2] Dept. of Computer Science & Engineering, Hanyang University, Ansan, Korea
Email: {joonywlee,sparky,scottlee}@hanyang.ac.kr

***Abstract*—Recent advancements in artificial intelligence (AI) have led to increased integration of AI technologies in various aspects of our lives. Deep learning systems, in particular, have proven effective in many areas. However, ensuring the quality, reliability, and safety of deep learning systems in critical environments is challenging. Traditional software testing methods are not suitable for deep learning systems. While there have been efforts to develop reliable testing techniques, there is a lack of focus on efficiency. As deep learning systems become larger and more complex, testing becomes more difficult and resource-intensive. The research community lacks a standardized efficiency metric for deep learning testing methods. We propose an efficiency metric that combines neuron coverage and test data accuracy to objectively evaluate the efficiency of white-box deep learning testing methods.**

Keywords—deep learning testing, efficiency metric, software testing

I. Introduction

The recent advancements in artificial intelligence technologies have greatly increased the integration of artificial intelligence into various aspects of our lives. Of the artificial intelligence technologies, deep learning systems have proven their effectiveness and efficiency in numerous areas. As the prevalence of deep learning systems increased, they were adopted into more areas including systems where a fault may result in critical consequences. In response to the critical environments the deep learning systems are put into, a method of ensuring these systems' quality, reliability, and safety became necessary. However, due to the nature of deep learning systems, traditional software testing methods are not effectively applicable to deep learning systems.

Both the research community and industry are striving to develop and implement techniques that offer reliable testing of deep learning systems, and, in turn, have achieved in coming up with various new and effective approaches for various areas such as medical diagnosis, autonomous driving, and computer vision. [5,6,7] Much research has been and is continuing to be done to further advance these approaches to achieve more promising results, but most of this research solely focuses on the effectiveness of the testing methods, not paying much attention to the efficiency of these methods.

As the size and complexity of problems that modern deep learning systems try to solve increase, the system itself is also increasing in size and complexity. The size and complexity of these deep learning systems make it more difficult to test and ensure the reliability of the systems, and also require more resources and/or time to test these systems. Additionally, as most testing processes of software development tend to consume a large part of the whole development process, the efficiency of the testing methods seems of great importance. However, the deep learning testing research community seems to be lacking a well-facilitated efficiency metric that can be utilized as a standard to further investigate the efficiency of a given method. Most approaches measure the efficiency of the testing methods naively using the actual time the testing method used up or the accuracy of the tests generated. This naive approach can be moderately indicative of the efficiency of testing tools and methods, but as the environment, the complexity of the problem, and the model size of deep learning systems can vary greatly, it cannot be properly utilized to objectively judge the efficiency.

We propose an efficiency metric that may allow testers and researchers alike to more objectively observe the efficiency of white-box deep learning testing methods utilizing both neuron coverage and accuracy of test data.

II. Background

A. Neuron Coverage

Traditional software testing utilized code coverage as a testing metric to observe the amount of logic explored by a test. However, most deep learning systems' logic is in the neurons of a neural network instead of the source code itself. Also, the logic of the system is mostly unexplainable even to its developer. Due to this nature, the coverage of deep learning logic is measured using neuron coverage. Neuron coverage is estimated by setting a threshold that determines whether a neuron is activated. If a neuron surpasses the set threshold during a test, it is considered activated and included as a covered neuron. Neuron coverage has proven its effectiveness as a metric that shows the quality of a test input of a deep learning system. [1] Although neuron coverage is indicative of the test quality and reliability, it is only available for white-box testing.

B. Deep Learning Test Input

Testing deep learning systems has introduced numerous new problems to the testing community. One of the most prominent challenges is the oracle problem. As deep learning systems are at most times non-deterministic there is not an easily accessible oracle to the test inputs. Also, as the logic of deep learning systems is embedded in the neural networks, not even the developer is able to understand the logic fully. To solve the challenges many new testing methods have been developed, many of which make perturbations to the test input or generate new test input with automated methods of labeling the data. [2,3,4] Also, as a metric most methods utilize neuron coverage.

III. Research Questions

To investigate the effectiveness of our proposal, we facilitate 3 research questions to investigate.

A. RQ1:Which parameter shows the most influence on the metric's results?

The efficiency metric of our proposal may output fluctuating results depending on the neuron activation threshold value and the model size. For the corresponding research question, we will investigate which of the factors, when parameterized, will show the greatest fluctuation in the results.

B. RQ2: Is there a range of values of parameters that seem most promising?

This research question searches for values of the parameters mentioned in the previous research question in hopes of finding the most optimal values. We define optimal as the value that achieves the most consistent results as the actual efficiency of the test.

C. RQ3: Does the proposed metric show consistency with actual efficiency?

This research question seeks validation of the proposed metric that proves the metric's results show consistency with the actual efficiency of the testing method. Consistency between the metric results and the actual efficiency can prove the efficiency metric to be a reliable metric to estimate the efficiency of a testing tool or method.

IV. Methodology

We plan to experiment and observe in order to answer our research questions and ultimately determine the reliability and validity of the proposed efficiency metric. To consider various environments in which deep learning may be utilized, we will use models of various fields including computer vision, speech recognition, and autonomous driving systems. The proposed efficiency metric will take into account the neuron coverage of an input test set and the accuracy, or the amount of error-inducing inputs the testing method generates. We may also give weighting values to either of the two factors depending on the results shown through experimentation.

A. RQ1:Influential Parameters

To answer this research question, we will run various test tools and methods on various models using different parameter values for the efficiency metrics. We can then observe the overall differences the changes to the parameter values make. The more differences made in the results means that the parameter shows greater influence in the metric's results.

B. RQ2: Optimal Parameter Values

To answer the optimal parameter value question, we plan to, similar to the method stated before, run various tests with varying values of parameters. We will record each parameter value and the corresponding consistency of efficiency. The calculation of consistency of efficiency will be elaborated in the next section.

C. RQ3: Consistency of Efficiency Metric

As the efficiency metric should provide a good indication of a testing method or tool's efficiency, we must investigate into the consistency between the results of the metric and the actual efficiency. We hope to be able to show consistency between the two by taking the time and resources taken by various test methods or tools on the same pre-trained model and also calculate the efficiency using the proposed metric. We can then take the differences of actual time and resources used between the various methods and compare the difference with the efficiency metric results. As the ratio of differences between the metric and the actual results come closer to a value of 1, it will indicate a stronger consistency between the actual efficiency and the metric results.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2023R1A2C1006390).

References

[1] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana."DeepXplore: automated whitebox testing of deep learning systems," Commun. ACM 62, 11 (November 2019), 137–145. https://doi.org/10.1145/3361566

[2] Boxi Yu, Zhiqing Zhong, Xinran Qin, Jiayi Yao, Yuancheng Wang, and Pinjia He, "Automated testing of image captioning systems," In proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2022). Association for Computing Machinery, New York, NY, USA, 2022, 467-479.

[3] Pin Ji, Yang Feng, Jia Liu, Zhihong Zhao, and Zhenyu Chen, "ASRTest: automated testing for deep-neural-network-driven speech recognition systems," In Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2022). Association for Computing Machinery, New York, NY, USA, 2022, 189–201. K. Elissa, "Title of paper if known," unpublished.

[4] Zixi Liu, Yang Feng, and Zhenyu Chen. 2021, "DialTest: automated testing for recurrent-neural-network-driven dialogue systems," In Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2021). Association for Computing Machinery, New York, NY, USA, 2021, 115–126. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] An Guo, Yang Feng, and Zhenyu Chen, "LiRTest: augmenting LiDAR point clouds for automated testing of autonomous driving systems," In Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2022). Association for Computing Machinery, New York, NY, USA, 2022, 480–492.

[6] Yuchi Tian, Ziyuan Zhong, Vincente Ordonez, Gail Kaiser,, and Baishaki Ray, "Testing DNN image classifiers for confusion & bias errors," In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20). Association for Computing Machinery, New York, NY, USA, 2020, 1122-1134.

[7] Yunhan Hou, Jiawei Liu, Daiwei Wang, Jiawei He, Chunrong Fang, and Zhenyu Chen, "TauMed: test augmentation of deep learning in medical diagnosis," In Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2021). Association for Computing Machinery, New York, NY, USA, 2021, 674-677.