Using Metamorphic Testing to Evaluate DNN Coverage Criteria

Jinyi Zhou, Kun Qiu, Zheng Zheng*
Beihang University
Beijing, China.
Email: {zy1803193, qiukun, zhengz}@buaa.edu.cn

Tsong Yueh Chen
Swinburne University of Technology
Melbourne, Australia.
Email: tychen@swin.edu.au

Pak-Lok Poon
Central Queensland University
Melbourne, Australia.
Email: p.poon@cqu.edu.au

I. Introduction

Generating test cases and further evaluating their "quality" are two critical topics in the area of Deep Neural Networks (DNNs). In this domain, different studies (e.g., [1], [2]) have reported that metamorphic testing (MT) serves as an effective test case generation method, where an initial set of source test cases is augmented with identified metamorphic relations (MRs) to produce the corresponding set of follow-up test cases. As a result, the fault detection effectiveness (and, hence, the "quality") of the resulting test suite T, containing these source and follow-up test cases, will most likely be increased.

Recently, we observed that some coverage criteria [3], [4] have been proposed to measure the quality of the test suites in the DNN domain. This observation leads to the following interesting and worth exploring research question (***RQ***):

Do these DNN coverage criteria properly reflect the quality improvement after a test suite has been augmented with MRs? We conducted a preliminary empirical study to answer ***RQ***.

II. Background Concepts

A. DNN Coverage Criteria

After an initial DNN classifier has been trained for some times, it should be tested whether or not its performance is satisfactory. In this regard, some coverage criteria have recently been proposed to determine the test adequacy of a trained classifier, with respect to the number of different run-time behaviors of a DNN that are triggered by a test suite.

In essence, a coverage criterion divides the output range obtained during the training of each neuron into smaller sub-ranges, and counts the neurons that reach those sub-ranges again in testing [3]. There are five recently proposed DNN coverage criteria, namely the k-multisection Neuron Coverage (KMNCov), the Neuron Boundary Coverage (NBCov), the Strong Neuron Activation Coverage (SNACov), the Top-k Neuron Coverage (TKNCov), and the Top-k Neuron Patterns (TKNPat). Their detailed discussions can be found in [3].

B. Metamorphic Testing (MT)

According to [1], applying MT to test a DNN classifier involves five steps as follows: (1) Manually identify a set of MRs, which are *expected properties* derived from the classifier's requirements or specifications. (2) Initially select some *source* test cases. (3) For each MR, generate some *follow-up* test cases from the relevant source test cases. (4) For each source test case t, execute the DNN classifier with t. Then, execute the classifier again with the follow-up test case t' corresponding to t. (5) Check whether or not the two execution results in (4) violate the relevant MR. If yes, a fault in the classifier is revealed and the classifier is labeled as unqualified. In the above, the test suite used for testing the classifier contains both the source and follow-up test cases.

III. Experiments

A. Experimental Setup

Subject DNN classifier. A handwriting digital recognition classifier (denoted by $\mathcal{C}$) was selected for our study. It was trained using LeNet-5 (one type of DNN model), and with a training dataset D collected from MNIST (a database containing labeled handwriting digital images).[1]

Metamorphic Relation (MRs). Here we briefly outline how MRs were used to improve the fault detection effectiveness of an original test suite for testing $\mathcal{C}$. Consider, for example, an image I with label $L(I)$, in D. We identified three MRs (MR_1^I, MR_2^I, MR_3^I) for $\mathcal{C}$. We use I' to denote the resulting image after changing I according to an MR.

- With respect to MR_1^I and MR_2^I, $\mathcal{C}$ should predicate the *same* label if I is slightly rotated or shifted by a small distance. Formally, if $I' = A_i \cdot I$, then $L(I') = L(I)$, where $i = 1$ or 2; A_1 and A_2 denote the rotation and translation matrices for MR_1^I and MR_2^I, respectively.
- With respect to MR_3^I, $\mathcal{C}$ should not be disturbed by the uncontrolled oscillation of hand muscles during writing [5]. Formally, if $I' = OS(I)$, then $L(I') = L(I)$, where $OS(\cdot)$ denotes the hand oscillation simulation function with respect to I (as discussed in [5]).

Experimental procedures. We used the five DNN coverage criteria mentioned in Section II.A to evaluate two types of test suites: one (T_r) generated by randomly selecting images from the test dataset of MNIST, and the other (T_m) generated by augmenting T_r with MRs. Assume $T_r = \{t_1, \cdots, t_{k/2}, t_{k/2+1}, \cdots, t_k\}$, then $T_m = \{t_1, \cdots, t_{k/2}, t'_1, \cdots, t'_{k/2}\}$, where k denotes the size a test

* Corresponding author.

[1] http://yann.lecun.com/exdb/mnist/

TABLE I
EXPERIMENTAL SETTING

Exp. index	MR Used	Size	Exp. index	MR Used	Size
1	None	500	2	None	1 000
3	MR_1	500	4	MR_1	1 000
5	MR_2	500	6	MR_2	1 000
7	MR_3	500	8	MR_3	1 000

suite; $|T_r| = |T_m| = k$; $t_i\,(1 \leq i \leq k)$ denotes a randomly selected image from the test dataset of MNIST; $t'_i\,(1 \leq i \leq k/2)$ denotes the follow-up test case generated for t_i according to one of the three MRs. Our study involved two values of k (500 and 1 000) and eight experiments as shown in TABLE I. In each of these eight experiments, due to "random" factors, 10 test suites were generated. Then, for each test suite T_r or T_m, we measured: (a) its prediction accuracy, and (b) its coverage score with respect to each of the five coverage criteria.

B. Experimental Results and Analysis

Fig. 1 summarizes the experimental results. Fig. 1(a) shows that those test suites (T_m's) generated by using MRs (Exp. indices 3 to 8) have smaller prediction accuracy (or simply accuracy) values, when compared with those test suites (T_r's) generated without using any MR (Exp. indices 1 and 2). For example, the accuracy scores of Exp. index 3 ($= 0.89 \pm 0.012$) were significantly smaller than those of Exp. index 1 ($= 0.99 \pm 0.004$), thereby indicating that T_m's detected more misbehaviors of $\mathcal{C}$ than T_r's. This observation is consistent with the results reported in other studies (e.g., [1], [2]), that MT generates higher-quality test suites for testing DNNs.

An examination of Figs. 1(b) to (f) provides an "initial" answer to ***RQ*** in three aspects. First, for NBCov and SNACov, they generally reflected the higher quality of T_m's generated in accordance with MR_1^I, MR_2^I, and MR_3^I. More specifically, in Figs. 1(c) and (d), the coverage scores of Exp. indices 3 to 8 (involving the use of MRs) were significantly larger than the corresponding scores of Exp. indices 1 and 2 (without using MRs). Second, similar to NBCov and SNACov, TKNCov and TKNPat generally reflected the higher quality of T_m's. More specifically, in Figs. 1(e) and (f), the coverage scores related to MR_2^I and MR_3^I were significantly larger than their corresponding scores without using any MR. However, the differences in these coverage scores between using MR_1^I and and not using MR_1^I were not as large as those differences between using and not using MR_2^I and MR_3^I. Third, for KMNCov in Fig. 1(b), it could not clearly distinguish between the quality of T_m's and T_r's (as the two lines in this figure are fairly horizontal), due to its method of dividing and counting covered neurons.

IV. CONCLUSION

Our study is the first to investigate the robustness of neuron coverage criteria from the perspective of MRs. We noted a worth taking observation that *only some* of these specific coverage criteria can effectively reflect the difference in quality between those test suites generated by using MRs and those are not. This observation suggests that more comprehensive studies should be conducted to further explore this issue, so that a better understanding between MT and DNN coverage criteria can be established.

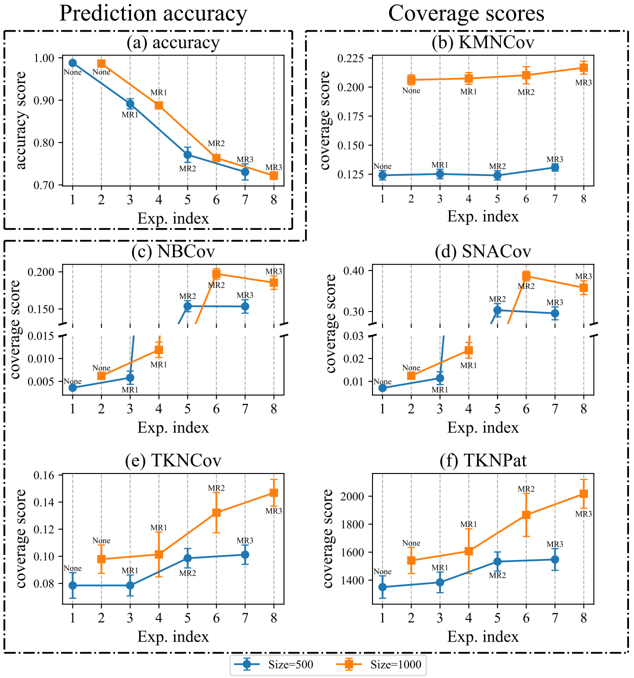

Fig. 1. (a) Prediction accuracy scores of $\mathcal{C}$ for the eight experiments. (b)–(e) Coverage scores of test suites with respect to the five coverage criteria. The x-axes correspond to the "Exp. index" columns of TABLE I. A dot represents the average value of accuracy scores (in Fig. (a)) or the average value of coverage scores (in Figs. (b)–(f)), with vertical bars representing the standard deviations.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61772055 and 61872169), the Technical Foundation Project of Ministry of Industry and Information Technology of China (Grant No. JSZL2016601B003), and the Equipment Preliminary R&D Project of China (Grant No. 41402020102).

REFERENCES

[1] A. Dwarakanath *et al.*, "Identifying implementation bugs in machine learning based image classifiers using metamorphic testing," in *Proceedings of 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2018, pp. 118–128.

[2] Z. Q. Zhou and L. Sun, "Metamorphic testing of driverless cars," *Communications of the ACM*, vol. 62, no. 3, pp. 61–67, 2019.

[3] L. Ma *et al.*, "DeepGauge: Multi-granularity testing criteria for deep learning systems," in *Proceedings of 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 120–131.

[4] K. Pei, Y. Cao, J. Yang, and S. Jana, "DeepXplore: Automated whitebox testing of deep learning systems," *Communications of the ACM*, vol. 62, no. 11, pp. 137–145, 2019.

[5] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proceedings of 7th International Conference on Document Analysis and Recognition*, 2003, pp. 958–963.