Testing Adequacy of Convolutional Neural Network Based on Mutation Testing

Yi Yao, Jialuo Liu*, Song Huang, Zhanwei Hui, Kaishun Wu, Lele Chen, Sen Yang,and Qiang Chen
Command and Control Engineering College
Army Engineering University of PLA
Nanjing, China
e-mail: 17551037134@163.com

***Abstract*—It is difficult to apply traditional testing adequacy criteria when measuring the adequacy of convolutional neural network applications. However, only a small number of test cases applied to the CNN model can achieve neuron coverage of almost 100%, overturning the effectiveness of the neuronal coverage criteria. In this paper, we propose a model coverage criterion based on mutation testing for CNN, and applying model coverage criterion to a common CNN image classification models (LeNet-5). we focus on the testing accuracy of model. Experiments show that our method can find the local optimal model and play an important role in improving the testing adequacy of the set of models.**

Keywords-CNN; testing adequacy; mutation testing; model coverage criterion

I. INTRODUCTION

We have witnessed great success of Convolutional Neural Network (CNN) in practical applications. All of the image classification and recognition, natural language processing have seen a near human level performance. Field of CNN applications will expand, and CNN would be also involved in many fields of related security. However, due to some faults in the CNN system, people are paying more and more attention to the security and reliability of CNN applications. To ensure the quality and security of CNN applications, DeepXplore [1] proposes a white box differential testing algorithm for systematically generating adversarial examples of all neurons in the overlay network and Anurag Dwarakanath applied Metamorphic testing to alleviate the testing oracle problem by CNN test oracle problem. For the testing adequacy of CNN, the existing methods and criteria of the testing adequacy of traditional software cannot be directly applied to the testing of CNN, because CNN has some properties as follows: Data sensitivity, Incomprehensibility, Parameterization of the program under test(PUT).

In this paper, in order to improve the testing adequacy of CNN, we propose the model coverage criterion based on mutation testing. The model coverage criterion can verify to a certain extent whether the design of the CNN model is reasonable. That is, whether the developer of the neural network selects a local optimal model. Compared to the neuron coverage criterion, The model coverage criterion, whose effectiveness is much higher in the test of the CNN model, has a stronger capability of testing adequacy of the CNN from the aspect of model design. We design six different types of mutation operators in the common CNN image classification model (LeNet-5) [3], and we focus on the testing accuracy of model. Experiments show that our method can find the local optimal model and play an important role in improving the testing adequacy of the set of models.

TABLE I. LENET-1,LENET-4,LENET-5 MODEL STRUCTURE

LeNet-1	LeNet-4	LeNet-5
Conv(4,5,5)+tanh	Conv(4,5,5)+tanh	Conv(6,5,5)+tanh/relu
MaxPooling(2,2)	MaxPooling(2,2)	MaxPooling(2,2)
Conv(12,5,5)+tanh	Conv(16,5,5)+tanh	Conv(16,5,5)+tanh/relu
MaxPooling(2,2)	MaxPooling(2,2)	MaxPooling(2,2)
Flatten()	Flatten()	Flatten()
FC(10)+Softmax	FC(120)+ tanh	Conv(120,5,5)+ tanh/relu
	FC(10)+Softmax	FC(84)+ tanh/relu
		FC(10)+Softmax

II. PROPOSED APPROACH

Yann LeCun [3] summarized three early convolutional neural network models, i.e., the LeNet-1, LeNet-4, and LeNet-5. The specific structure is shown in the following table.From Table I, by changing the number of convolution kernels, the number of fully connected layers, the number of convolution layers, and the activation function, we can change LeNet-1 to LeNet-4, LeNet-5. Inspired by this method, we have summarized six types of mutation operators (changing the number of convolution kernels, the number of fully connected layers, the number of convolution layers, the activation function, the manner of padding and the size of the convolution kernel). For a CNN model M, if the mutation model $\{M_1', M_2', \dots, M_n'\}$ obtained by injecting the above six mutation operators into the original model M. We treat $\{M, M_1', M_2', \dots, M_n'\}$as the same type of model. Because the final model is based on this type of model by adjusting and optimizing. The type of CNN model named a mutation model class is regarded as the object to be tested. By improving the coverage of the objects to be tested, we can improves the testing adequacy of this type of model to some extent. For traditional software, we combine many testing methods (equivalent class, boundary value, metamorphic test) to construct the reliable test suite.

* Corresponding author.

978-1-7281-3925-8/19/$31.00 ©2019 IEEE
DOI 10.1109/QRS-C.2019.00112

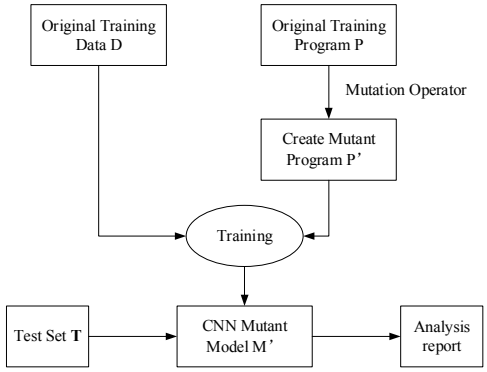

Figure 1. The general process of CNN model mutation testing

By improving the code coverage of the software code, branch coverage and condition coverage, we can improve testing adequacy. For the CNN model, a few test cases are likely to cover the entire model. Therefore, for the CNN model, it is impossible to have the reliable test suite. According to the existing the theory of reliable test suite, we propose the concept of the set of mutation model. If the mutation model set is adequate for the mutation operators, the set of mutation model is called the set of reliable mutation models. In this way, the testing adequacy of each target CNN model can be improved based on the set of reliable mutation models.

Definition of the mutation model class: By injecting all kinds of mutation operators into the original program P, we can get many mutation programs $\{P_1', P_2', \dots, P_n'\}$. Then the relevant data sets are trained in these models, you can get a series of completed training mutation models $\{M_1', M_2', \dots, M_n'\}$. $\{M, M_1', M_2', \dots, M_n'\}$ is treated as the same model.

Based on the above analysis, we propose the definition of model coverage: a measure used to describe the degree to how many models in the set of CNN models is tested.

III. EXPERIMENTS

In this section, We designed six mutation operators and applied them to common CNN applications to verify the effectiveness of the model coverage criteria.

The general process of variation testing based on convolutional neural network software is shown in Figure 1. The tester injects different mutation operators into the original convolutional neural network training program, and can obtain the corresponding mutated CNN program $\{P_1', P_2', \dots, P_n'\}$. Then, the test set T of the original model M is executed on each mutation program $\{P_1', P_2', \dots, P_n'\}$ to obtain the corresponding mutation model $\{M_1', M_2', \dots, M_n'\}$. The more mutation models obtained using this method, the higher the testing adequacy. At the same time, we can compare the testing adequacy of all models and select the model with the highest accuracy. The model is named as a local optimal model.

We have set up nine mutants for the CNN model as follows: Change Activation Function, Change the manner of padding, Reduce Convolution Layers, Add Convolution Kernels, Reduce Convolution Kernels, Enlarge Convolution Kernel Size, Cut Convolution Kernel Size, Add Fully Connected Layers, Reduce Fully Connected Layers. Injecting mutations into the original program P will directly change the convolutional neural network model. Since the convolutional neural network model may be affected by many parameters. We screened nine different mutants that would directly alter the structure of the convolutional neural network.

TABLE II. THE ACCURACY OF EACH MODEL

Model	Test Accuracy
Original	98.06%
Change Activation Function (CAF)	98.00%
Change the Manner of Padding (CMP)	98.01%
Reduce Convolution Layers (RCL)	96.26%
Add Convolution Kernels (ACK)	98.65%
Reduce Convolution Kernels (RCK)	97.65%
Enlarge Convolution Kernel Size (ECKS)	98.00%
Cut Convolution Kernel Size (CCKS)	98.20%
Add Fully Connected Layers (AFCL)	99.03%
Reduce Fully Connected Layers (RFCL)	97.06%

The mutation testing criteria are usually measured by the mutant score. For the mutation CNN model, any test case $t \in T$ can be correctly classified by the original CNN model M, but cannot be correctly classified by the variant CNN model M'. The test case t kills the mutation M'. The mutant score refers to $mutants_{killed}/mutants_{all}$. However, the mutant score of traditional software does not apply to the CNN, due to the large number of the testing set T, it is very easy to kill the mutation M' for the test case $t \in T$. For the above reasons, for the mutation testing of CNN, we inject the mutation operators into the CNN training program. Retrain the executive program using the set of training data to generate the corresponding mutation CNN model $\{M, M_1', M_2', \dots, M_n'\}$. The more mutation models obtained using this method, the higher the Testing adequacy . At the same time, we can compare the testing accuracy of all models and select the model with the highest accuracy. The model is named as local optimal model.

IV. PRELIMINARY RESULTS AND CONCLUSIONS

We concluded three characteristics of CNN and took advantage of the mutation testing in CNN and proposed a mutation framework and six mutation operators based on CNN, we also proposed a model coverage criteria at the same time. Experiments showed that our method is beneficial to improve the testing adequacy of the model, so that the quality and safety of CNN applications could be guaranteed.

ACKNOWLEDGMENT

This work is supported by National Key R&D Program of China(No:2018YFB1403400).

REFERENCES

[1] K. Pei, Y. Cao, J. Yang, and S. Jana, "DeepXplore: Automated Whitebox Testing of Deep Learning Systems," pp. 1-18, 2017.

[2] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.