Evaluation of T-Way Testing of DNNs in Autonomous Driving Systems

Jaganmohan Chandrasekaran, Ankita Ramjibhai Patel
Department of Computer Science & Engineering
The University of Texas at Arlington
Arlington, USA
{jaganmohan.chandrasekaran, ankitaramjibhai.patel}@mavs.uta.edu

Yu Lei
Department of Computer Science & Engineering
The University of Texas at Arlington
Arlington, USA
ylei@cse.uta.edu

Raghu Kacker, D. Richard Kuhn
Information Technology Lab
National Institute of Standards and Technology
Gaithersburg, USA
{raghu.kacker, d.kuhn}@nist.gov

***Abstract*—A Deep Neural Network (DNN) model is used to perform intelligent, safety-critical tasks in Autonomous Driving Systems (ADS). In our prior work, we proposed a combinatorial testing approach to test DNN models used to predict a car's steering angle. We generate test images by applying a set of combinations of basic image transformations. In this paper, we report a preliminary study that compares the performance of synthetic images generated using a combinatorial approach to DeepTest, a state-of-the-art tool that aims at generating test inputs that maximize neuron coverage. We present an experimental evaluation by measuring and comparing the neuron coverage achieved using the two approaches. Two pre-trained DNN models from the Udacity driving challenge are used as the subject DNNs. The results suggest that the combinatorial approach performs better than the DeepTest approach in generating valid synthetic images and covering an additional number of neurons.**

Keywords— Combinatorial Testing, DeepTest, Neuron Coverage.

I. Introduction

Deep Neural Network (DNN) models are used in autonomous driving systems to perform tasks such as pedestrian detection, steering control, object detection. Despite its promising potential, when applied in real-world conditions, the DNN models exhibit erroneous behavior resulting in life-threatening consequences [2]. It is vital to rigorously test these models before their deployment in the real world.

Our earlier work presented a combinatorial approach to generate synthetic images to test the pre-trained DNN models used in self-driving cars [1]. This paper reports two significant extensions of our earlier work. First, in addition to the neuron coverage results reported in [1], we report the neuron coverage for the Chauffeur model. Second, we present a comparative evaluation where we compare the neuron coverage results achieved by our approach to those achieved by DeepTest, a test generation approach that aims at generating test inputs that maximize the neuron coverage [4]. Neuron coverage is a measure of the proportion of neurons activated in a DNN model. Experimental results suggest that in most cases, t-way synthetic images cover an additional number of neurons compared to the DeepTest approach. The remainder of the paper is organized as follows. Section II presents a brief introduction to the t-way testing of DNNs. Section III presents the experimental design, results, and discussion. In Section IV, we present the concluding remarks and directions for future work.

II. T-Way Testing of DNNs

We presented a combinatorial approach to generate t-way synthetic images to test DNN models[1]. In this approach, First, we identify a set of valid image transformations applicable to the seed image. Next, we design an input parameter model (IPM) based on the valid transformations; each valid transformation is mapped as a parameter in the IPM. Then, based on the IPM, we generate an abstract t-way (t=2) test set. Each t-way test represents a combination of image transformations. Finally, using an image processing library, we generate synthetic images by applying the t-way image transformations to the seed image. The t-way synthetic images are used to test the DNN models.

III. Experiments

A. Experimental Design

Tian et al. evaluated the impact of synthetic images generated by combining different image transformations on the neuron coverage using three open-source DNN models, namely Rambo, Chauffeur, and Epoch[4]. In the case of Epoch, a pre-trained model is not publicly available for download. Therefore, we used the remaining two models, namely Rambo and Chauffeur, in this comparison study.

In their evaluation, they generated synthetic images using two approaches, namely *Cumulative transformations* and *Guided transformation*. Similar to our earlier work [1], the guided transformation approach generates synthetic images by combining a set of image transformations. However, this approach aims to generate tests that maximize the neuron coverage and does not guarantee to generate valid synthetic images. That is, while the synthetic images generated using the guided transformation approach can cover an additional set of neurons, they may not be used to determine the correctness of a DNN model because invalid images may never exist in reality.

Therefore, we compare the cumulative neuron coverage achieved by t-way synthetic images to those synthetic images generated using the cumulative transformation approach. We will refer to the cumulative transformation approach as the *DeepTest* approach unless otherwise specified.

To generate synthetic images using the DeepTest approach, we apply a set of valid image transformations identified for the respective seed image. We observed that in most cases, the number of synthetic images generated using a t-way test set is substantially higher compared to that of the DeepTest approach. Therefore, to facilitate a fair comparison, for each group, using a random sampling approach, we select a subset from the t-way test set (synthetic images) such that the number of the t-way tests in the subset is equal to the total number of synthetic images generated using the DeepTest approach.

Then, we execute the DNN model with the seed image (baseline), followed by t-way synthetic images from the subset, and measure the cumulative neuron coverage. We compare the cumulative neuron coverage achieved by the t-way subset with the synthetic images generated using the DeepTest approach. To reduce variations in random sampling, we generated five samples for each group by using different seeds (selected at random).

We refer the reader to our earlier work [1] for additional information about the measurement of cumulative neuron coverage, the number of seed images, the number of valid transformations, and the number of t-way test cases generated for each seed image.

B. Results and Discussion

First, we present the cumulative neuron coverage achieved by t-way tests for *Chauffeur*. The Chauffeur model consists of 1 CNN sub-model with 1427 neurons and 1 LSTM sub-model with 513 neurons. Tian et al. did not include the LSTM sub-model in their evaluation. Hence, for *Chauffeur*, we limit our comparison to the CNN sub-model.

For the Chauffeur model, 14 out of 19 seed images cover less than 15% of the total neurons (1427 neurons); Among the seed images, Group 7 covers the least, covering 6% of total neurons (90 neurons), while the seed image from Group 16 covers the most with 22% of total neurons (318 neurons).

Figure 1 presents the cumulative neuron coverage achieved by t-way tests for *Chauffeur.* The x-axis represents the group number. The y-axis represents the percentage of additional neurons covered by the t-way tests compared to their respective baseline. Our results suggest that t-way tests result in a significant increase in neuron coverage. Out of nineteen groups, t-way tests generated for sixteen groups achieve more than one hundred percent increase in cumulative neuron coverage.

Next, we present the comparison results. For *Rambo*, the coverage results obtained from our earlier work are re-used in our comparison experiments. Figure 2 and Figure 3 present the comparison results for *Rambo* and *Chauffeur,* respectively. The x-axis represents the group number. The y-axis represents the number of neurons. Due to space limitations, we present the average cumulative neuron coverage achieved by the five t-way subsets for each group. A horizontal blue bar in the bar chart indicates the cumulative neuron coverage achieved using the DeepTest approach. Our results indicate that for *Rambo*, in most cases (18 out of 19 groups), subsets of the t-way test set achieve a higher cumulative coverage compared to the DeepTest approach. For five groups (Group 2, 7, 8, 9, 13), the subset (of the t-way test) covers a significant number of additional neurons compared to the DeepTest approach.

In the case of *Chauffeur*, for 16 groups, all five samples of the t-way subset cover a significant additional number of neurons compared to the DeepTest approach. For the remaining three groups (Group 4, 5, 20), the t-way subset covers a marginally higher number of neurons than the DeepTest approach.

Overall, the results from this initial study indicate that synthetic images generated using the combinatorial approach can achieve higher neuron coverage than the DeepTest approach. The source code, results, data and/or artifacts have been made available at [3].

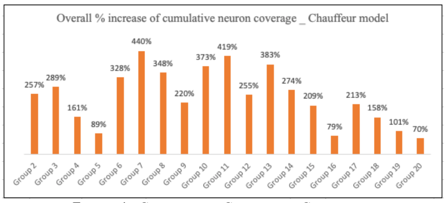

FIGURE 1 - CUMULATIVE COVERAGE - CHAUFFEUR

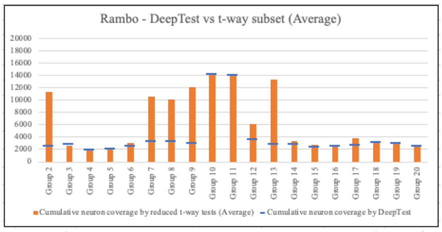

FIGURE 2 – COMPARISON – DEEPTEST VS. REDUCED T-WAY (RAMBO)

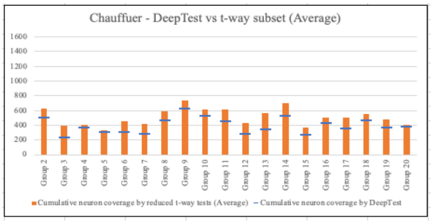

FIGURE 3 – COMPARISON – DEEPTEST VS. REDUCED T-WAY (CHAUFFEUR)

IV. CONCLUSION AND FUTURE WORK

In this paper, we present the cumulative neuron coverage for the Chauffeur model and an initial study that compares the synthetic images generated using a combinatorial approach to that of DeepTest in terms of cumulative neuron coverage. In most cases, the results suggest that the synthetic images generated using the combinatorial approach cover an additional number of neurons compared to the DeepTest approach.

As part of future work, we plan to conduct a comprehensive empirical study that compares the effectiveness of combinatorial testing to that of random testing in testing pre-trained DNN models.

ACKNOWLEDGMENT

This work is supported by research grant (70NANB18H207) from Information Technology Lab of National Institute of Standards and Technology (NIST).

Disclaimer: Certain software products are identified in this document. Such identification does not imply recommendation by the NIST, nor does it imply that the products identified are necessarily the best available for the purpose.

REFERENCES

[1] Chandrasekaran, J., Lei, Y., Kacker, R., & Kuhn, D. R. (2021, April). A Combinatorial Approach to Testing Deep Neural Network-based Autonomous Driving Systems. In *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 57-66). IEEE.

[2] Tesla driver dies in first fatal crash while using autopilot mode: https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk, Accessed: 2021-07-10.

[3] Testing-AI-Systems, https://tinyurl.com/k4aswyef

[4] Tian, Y., Pei, K., Jana, S., & Ray, B. (2018, May). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering* (pp. 303-314).