|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| P | **Methodology** | **Dataset** | **Benchmark** | **Limitation/Future Work** | **Coverage Criteria** | **Test Generation** |
| 1. | Symbolic execution with local explainability.  LIME provides explanations for predictions | * German Credit Data * Adult census income * Bank marketing * US Executions * Fraud Detection * Raw Car Rentals * credit data * census data | THEMIS (Algorithm)  The technique produces 3.72 times more successful test cases than existing state-of-the-art. | * FW: Expand to text and image domains * FW: Measure symbolic execution efficacy using neuron coverage, boundary value coverage. | -- | symbolic/concolic |
| 2 | concolic testing method | * MNIST * CIFAR-10 | DeepXplore DeepTest, DeepCover, and DeepGauge |  | NC, SSC, NBC | concolic |
| 3 | whitebox framework for testing DL systems, introducing neuron coverage for test measurement | * MNIST * ImageNet * Driving * VirusTotal * Drebin | LeNet variations  State-of-the-art image classifiers from ILSVRC  Nvidia DAVE self-driving systems  PDF malware detectors  Android app malware detectors | L: Inherits differential testing constraints. | NC | dual-optimisation |
| 4 | automates testing for DNN-driven autonomous cars. | Udacity self driving car challenge 2 | Chauffeur-  Epoch  Rambo-S1  Rambo-S2  Rambo-S3 | L: generates synthetic images via transformations on seed images, not designed to be exhaustive, possibly missing some realistic cases.  L: restricted only the accuracy of the steering angle , not focus on brake and accelerator  L: cannot simulate complex road scene transformations (e.g., snowy scenes). | NC | greedy search |
| 5 | White box methodology, Proposed four novel test criteria tailored to DNN ,structural features. able to capture and quantify causal relations existing in a DNN , Achieved balance between bug finding ability and computational cost. | MNIST , CIFAR-10 , ImageNet | state-of-the-art neural networks of different sizes (from a few hundred up to millions of neurons) to demonstrate their utility with respect to four aspects: À *bug finding* Á *DNN safety statistics* Â *testing efficiency* Ã *DNN internal structure analysis* |  | MC/DC | symbolic execution |
| 6 | proposed criteria facilitate the understanding of DNNs as well as the test data quality from different levels and angles. | MNIST, ImageNet | LeNet-1 LeNet-4 LeNet-5 ,pre- trained VGG-19, ResNet-50 | * Investigate DeepGauge on more diverse datasets and DL architectures. * Explore the application of DeepGauge in real-world DL systems. | NBC | gradient descent methods |
| 7 | an unsupervised learning framework to synthesize realistic driving scenes to test inconsis- tent behaviors | Udacity Training Udacity Test Ep1 Udacity Test Ep2 Youtube Ep1 Youtube Ep2 | Autumn, Chauffeur | lack a good stan- dard to evaluate image quality (i.e. realisticity)  Udacity dataset is relative small and the autonomous driving models are quite simple  Only focus on steering wheel | -- | Mutation testing |
| 8 | an auto- mated fuzz testing framework for hunting potential defects of general-purpose DNNs. | MNIST, CIFAR-10, ImageNet | LeNet-1  LeNet-4  LeNet-5  RN-20  VGG-16  MobileNet  RN-50 | In this paper, NC cannot generate effective results to evaluate the models with various quality  NC is less effective in error triggering test detection and sensitive defect detection. | Neuron Cov. (NC) K-multisec. Neu. Cov. (KMNC) Neuron Bound. Cov. (NBC) Strong Neuron Act. Cov. (SNAC) Top-k Neu. Cov. (TKNC) Bottom-k Neu. Cov. (BKNC) | Metamorphic mutation |

1. Automated Test Generation to Detect Individual Discrimination in AI Models.pdf
2. DeepConcolic
3. DeepXplore
4. DeepTest
5. DeepCover
6. DeepGauge
7. DeppRoad
8. DeepHunter