Adversarial Image Generation based on Various Neuron Coverage

Team8

20170181 Taeyoung Kim
20180650 Hyunjoon Cho
20205424 Sunjae Kwon

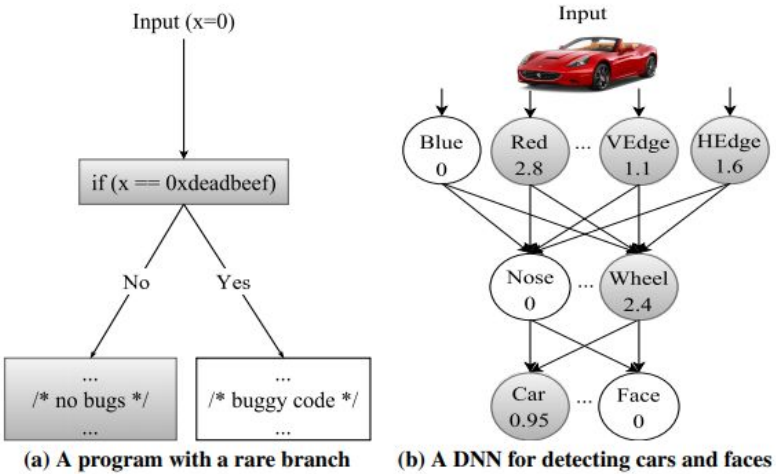
Introduction

- DNN(Deep neural networks) is widely applied to safety-critical applications
- Demand for testing validating the DNN is increasing
- However, it’s not easy to manually find rare input generating erroneous behavior.
- Thus, various **automated testing tools for DNN** has been studied.

Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." proceedings of the 26th Symposium on Operating Systems Principles. 2017.

Introduction

- **Code coverage & Neuron coverage**
  - Test cases having **higher code coverage** tend to **reveal fault** in the code
  - Test inputs having **higher neuron coverage** tend to occur **erroneous behavior**



(a) A program with a rare branch (b) A DNN for detecting cars and faces

Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." proceedings of the 26th Symposium on Operating Systems Principles. 2017.

Related works

- **DeepXplore**
  - **Objective** : Joint optimization of neuron coverage and differences in the prediction of DNN models
  - **Maximizing objective** generates test that achieve high neuron coverage while simultaneously achieve erroneous prediction

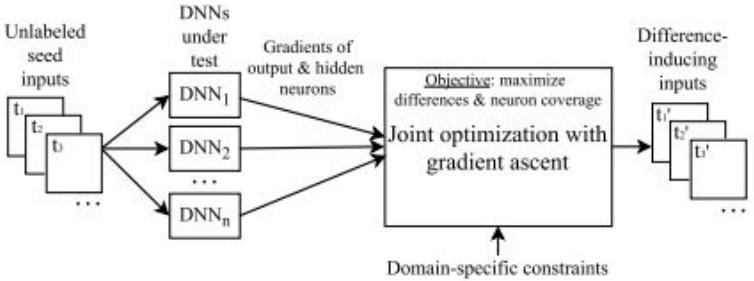


Figure 5: DeepXplore workflow.

Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." proceedings of the 26th Symposium on Operating Systems Principles. 2017.

Related works

- **DLFuzz**
  - Keeps minutely mutating the input to maximize the neuron coverage and the prediction difference between original input and the mutated input
  - d**oes not require multiple DNN models**

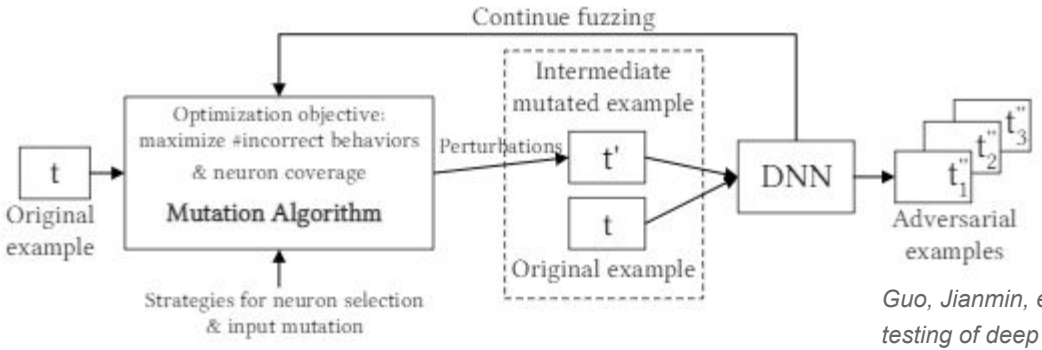


Figure 2: Architecture of DLFuzz

Guo, Jianmin, et al. "DLFuzz: differential fuzzing testing of deep learning systems." Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2018.

Neuron Coverage

- Given threshold ε, if neuron's activation value is larger than ε for all test input, it is covered.
- Proposed in DeepXplore.

Fig from Nils Wenzler. "Not all Neurons are created equal: Towards a feature level Deep Neural Network Test Coverage Metric" In Proceedings of . ACM, New York, NY, USA, 8 pages. 2019.

k-multisection Neuron Coverage

- Let h, l be highest neuron activation value from training data set (lowest, respectively).
- Range [l, h] is divided into k section.
- Coverage is measured with percentage of sections that at least one activation value was observed.
- Proposed in Deepgauge.

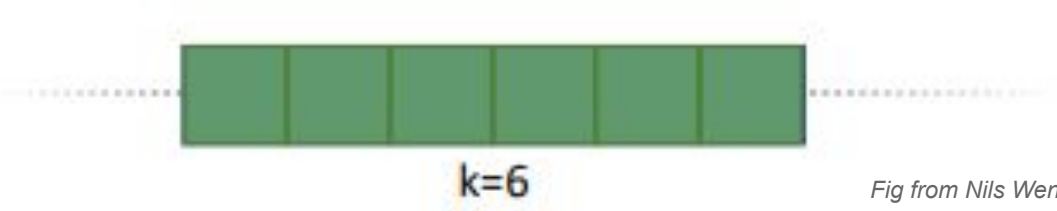


Fig from Nils Wenzler. "Not all Neurons are created equal: Towards a feature level Deep Neural Network Test Coverage Metric" In Proceedings of . ACM, New York, NY, USA, 8 pages. 2019.

Neuron Boundary Coverage

- Let h, l be highest neuron activation value from training data set (lowest, respectively).
- The neuron is covered if both upper corner region [h, ∞] and lower corner region [-∞, l] is covered.
- Corresponds boundary coverage in software testing.
- Proposed in Deepgauge.

Fig from Nils Wenzler. "Not all Neurons are created equal: Towards a feature level Deep Neural Network Test Coverage Metric" In Proceedings of . ACM, New York, NY, USA, 8 pages. 2019.

Strong Neuron Activation Coverage

- Very similar to Neuron Boundary Coverage, however only considers upper corner region.
- Proposed in Deepgauge.

Fig from Nils Wenzler. "Not all Neurons are created equal: Towards a feature level Deep Neural Network Test Coverage Metric" In Proceedings of . ACM, New York, NY, USA, 8 pages. 2019.

Other coverages

- Top-k neuron coverage (pattern) : Only considers neurons with top-k activation value, and their pattern.
- Sign/Value-Sign/Value Coverage : Sign change is observed if all features' sign are different for two input pair, value change when differ with significant difference.
- Surprise Adequacy : Activation value's distance/Likelihood to training dataset's activation values.

Problem Statement

- Various neuron coverage metric have been proposed

  ex) k-multisection Neuron Coverage, Neuron Boundary Coverage..

- However, **DeepXplore** and **DLFuzz** only consider a neuron coverage
- Need for considering these various neuron coverage

Project Goal

- Find which coverage works **best** in creation of adversarial input
  - More on Evaluation slide

Methodology

- Apply various coverage to DeepXplore/DLFuzz
  - Define gradient for each coverage

```
30: procedure COMPUTE_OBJ2(x, dnns, cov_tracker)
31:    loss := 0
32:    for dnn ∈ dnns do
33:           select a neuron n inactivated so far using cov_tracker
34:           loss += n(x) //the neuron n's output when x is the dnn's input
35:    return loss
```

Pseudo Code from Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." proceedings of the 26th Symposium on Operating Systems Principles. 2017.

Datasets & Target DNN models

- MNIST and LeNet variations
- ImageNet and VGG variations

Evaluation metrics

- Average neuron coverage improvement
- Number of adversarial inputs generated
- Average time of generating per adversarial input
    - Used in DLFuzz, to compare with DeepXplore