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 and validating the DNN is increasing
- However, it’s not possible to find all inputs generating erroneous behavior.
- Thus, various **automated testing tools for DNN** has been studied.

  ex) **DeepXplore** and **DLFuzz**

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

Problem Statement

- **DeepXplore** and **DLFuzz** depends on basic neuron coverage
- In the meantime, various neuron coverage metric have been proposed

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

- Need for considering these various neuron coverage

Project Goal

- Find which coverage works **best** in creation of adversarial input
- Analyze characteristics of each neuron coverage

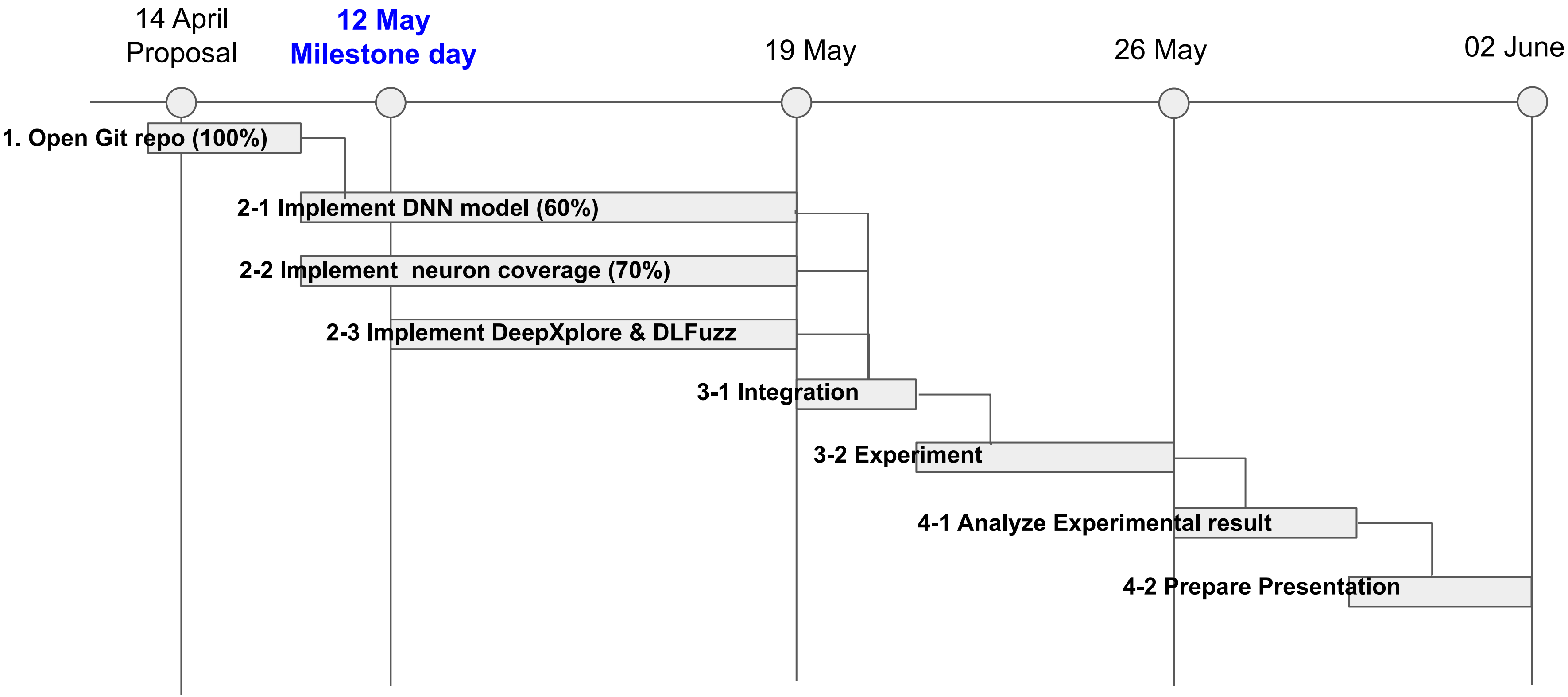
Coverage Analysis

- Neuron Coverage
  - Basic coverage, used in DeepXplore and DLFuzz
- k-multisection Neuron Coverage
  - Section is bounded by low/high output from training
  - Sections are already covered by training data, less probable to show new behavior
- (Strong-) Neuron Boundary Coverage
  - Increased coverage may invoke more logic, resulting in unexpected behavior
  - Neurons are activated by output over threshold, thus upper bound would be more influential

Current State

- Currently Done
  - Implement loading pretrained model VGGNet for ImageNet Dataset
  - Implement coverage computation tool
    - Neuron Coverage
    - k-multisection Coverage
    - Neuron Bounday Coverage
    - Strong Neuron Boundary Coverage
- Need to be done
  - Extracting activation value for uncovered neuron
  - Defining loss for each coverage with activation value
  - Implementing DeepXPlore, DLFuzz

TimeLine


14 April
Proposal
12 May
Milestone day
19 May
26 May
02 June
1. Open Git repo (100%)
2-1 Implement DNN model (60%)
2-2 Implement neuron coverage (70%)
2-3 Implement DeepXplore & DLFuzz
3-1 Integration
3-2 Experiment
4-1 Analyze Experimental result
4-2 Prepare Presentation