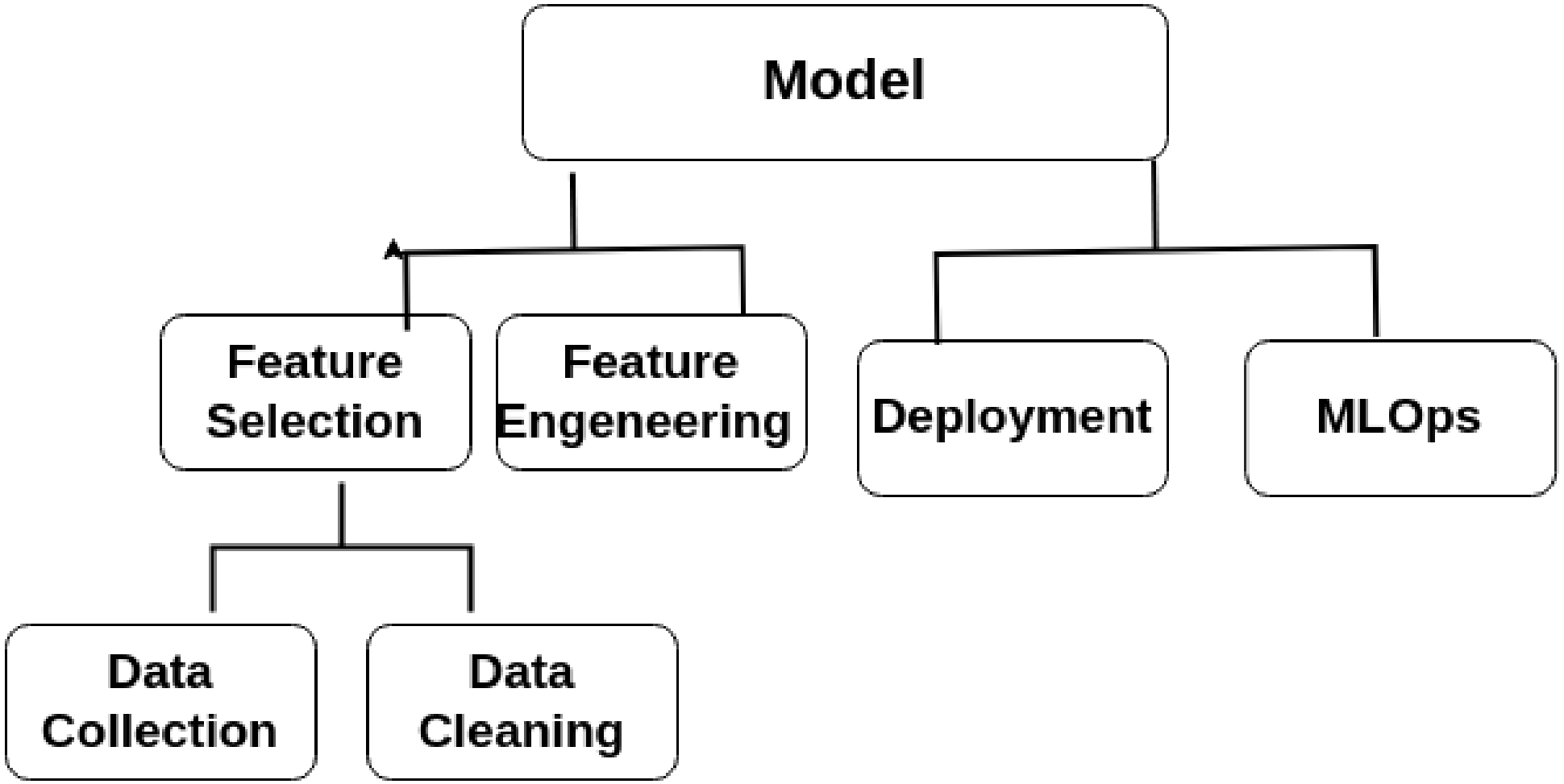
Reversing Deep Learning Models

Everyone Works For Me



Trine - Trine

Courtesy : WB

Models

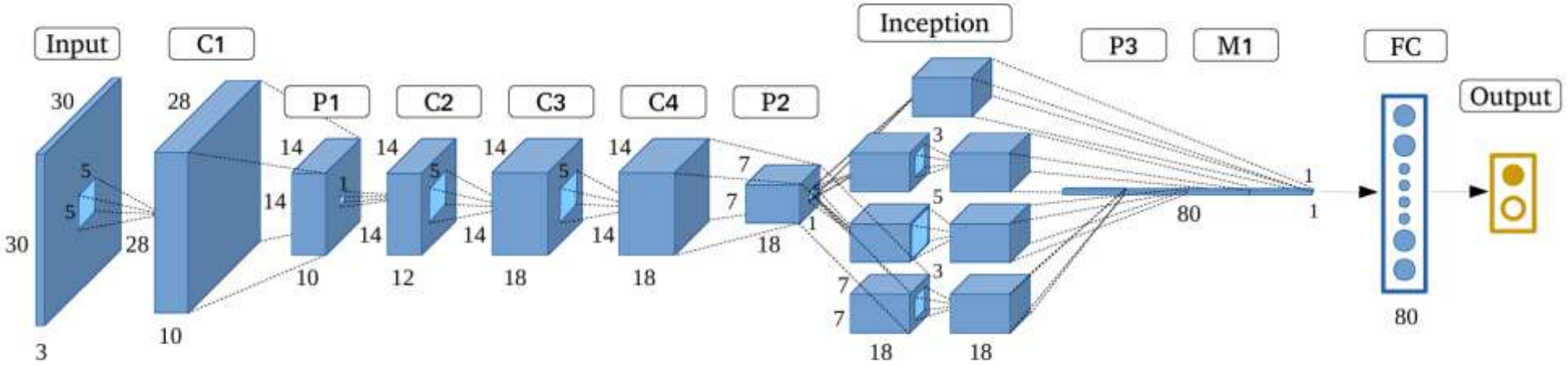
ONNX

Safetensors
ML Safer For All

TensorFlow Lite

Googlenet

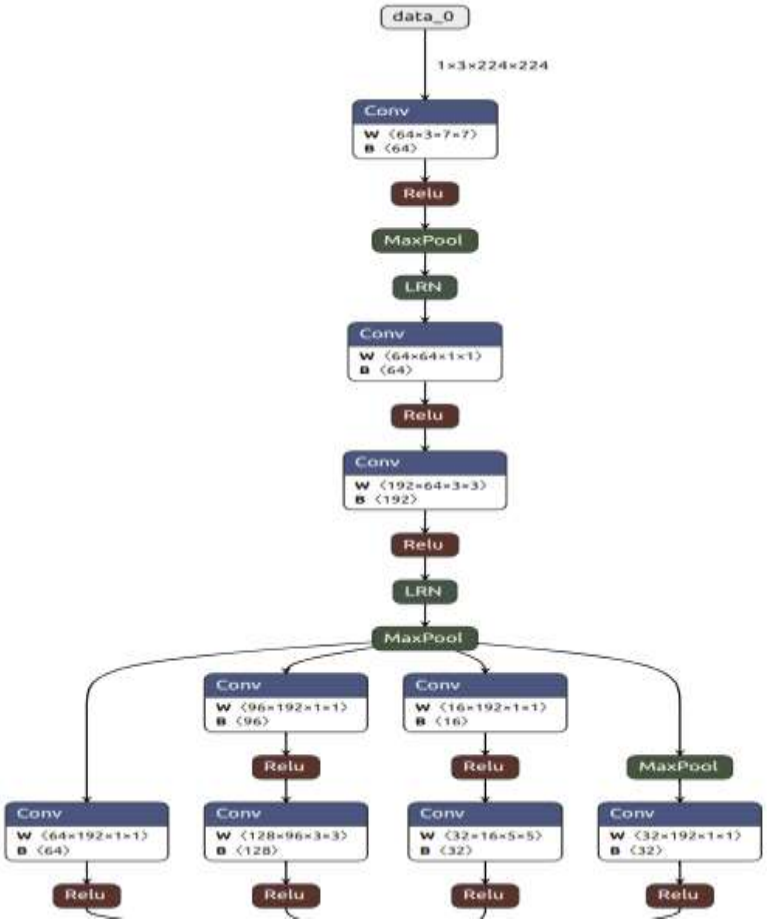

INCEPTION


Input
C1
P1
C2
C3
C4
P2
Inception
P3
M1
FC
Output
30
28
5
3
10
14
1
12
18
7
80

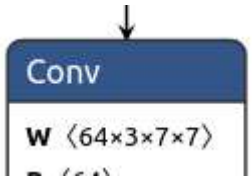
Reversing Models – Rugged Approach

```
graph main_graph (
  %x[FLOAT, 1x3x299x299]
) initializers (
  %fc.weight[FLOAT, 1000x2048]
  %fc.bias[FLOAT, 1000]
  %onnx::Conv_877[FLOAT, 32x3x3x3]
  %onnx::Conv_878[FLOAT, 32]
  %onnx::Conv_880[FLOAT, 32x32x3x3]
  %onnx::Conv_881[FLOAT, 32]
  %onnx::Conv_883[FLOAT, 64x32x3x3]
  %onnx::Conv_884[FLOAT, 64]
  %onnx::Conv_886[FLOAT, 80x64x1x1]
  %onnx::Conv_887[FLOAT, 80]
  %onnx::Conv_889[FLOAT, 192x80x3x3]
  %onnx::Conv_890[FLOAT, 192]
  %onnx::Conv_892[FLOAT, 64x192x1x1]
```

Netron- Breaking Confidentiality



How much I Weigh- Beyond Keys and Buffer Overflow

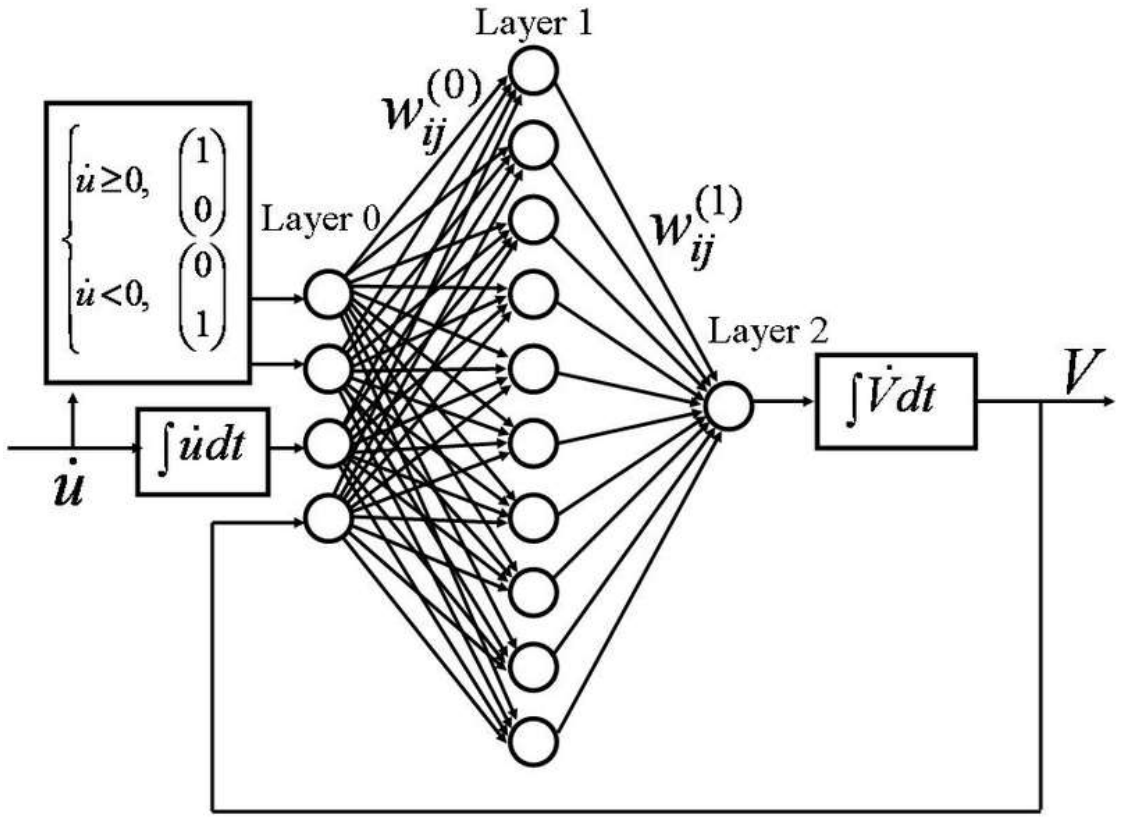


TENSOR PROPERTIES

name	conv1/7x7_s2_w_0
category	Initializer
type	float32
shape	64,3,7,7
value	

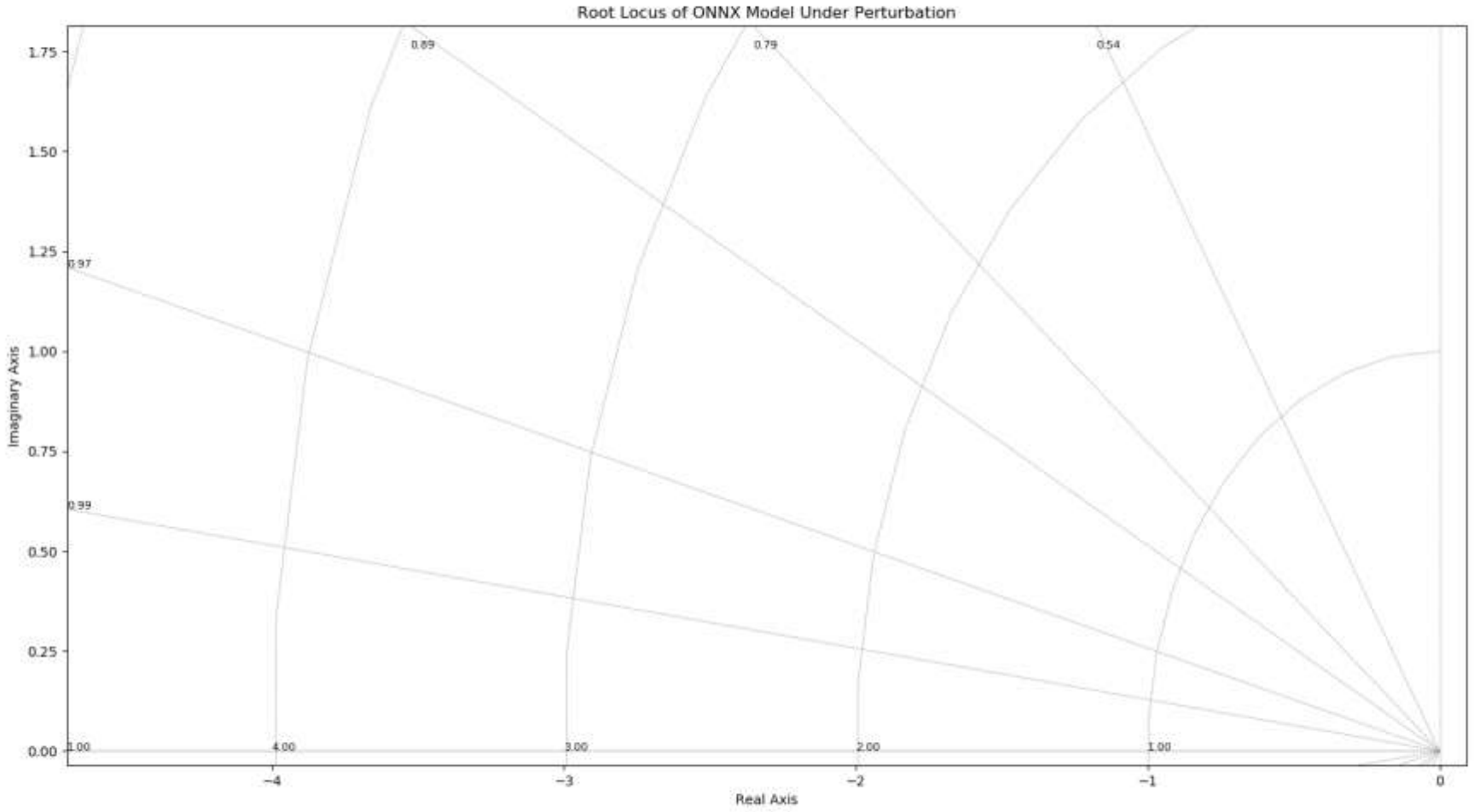
```
[
    [
        [
            [
                0.08820479363203049,
                0.03741685673594475,
                0.04910733550786972,
                0.16268368065357208,
                0.15742811560630798,
                0.017702095210552216,
                -0.003900340758264065
            ],
            [
```

Domain Jump – Control Systems

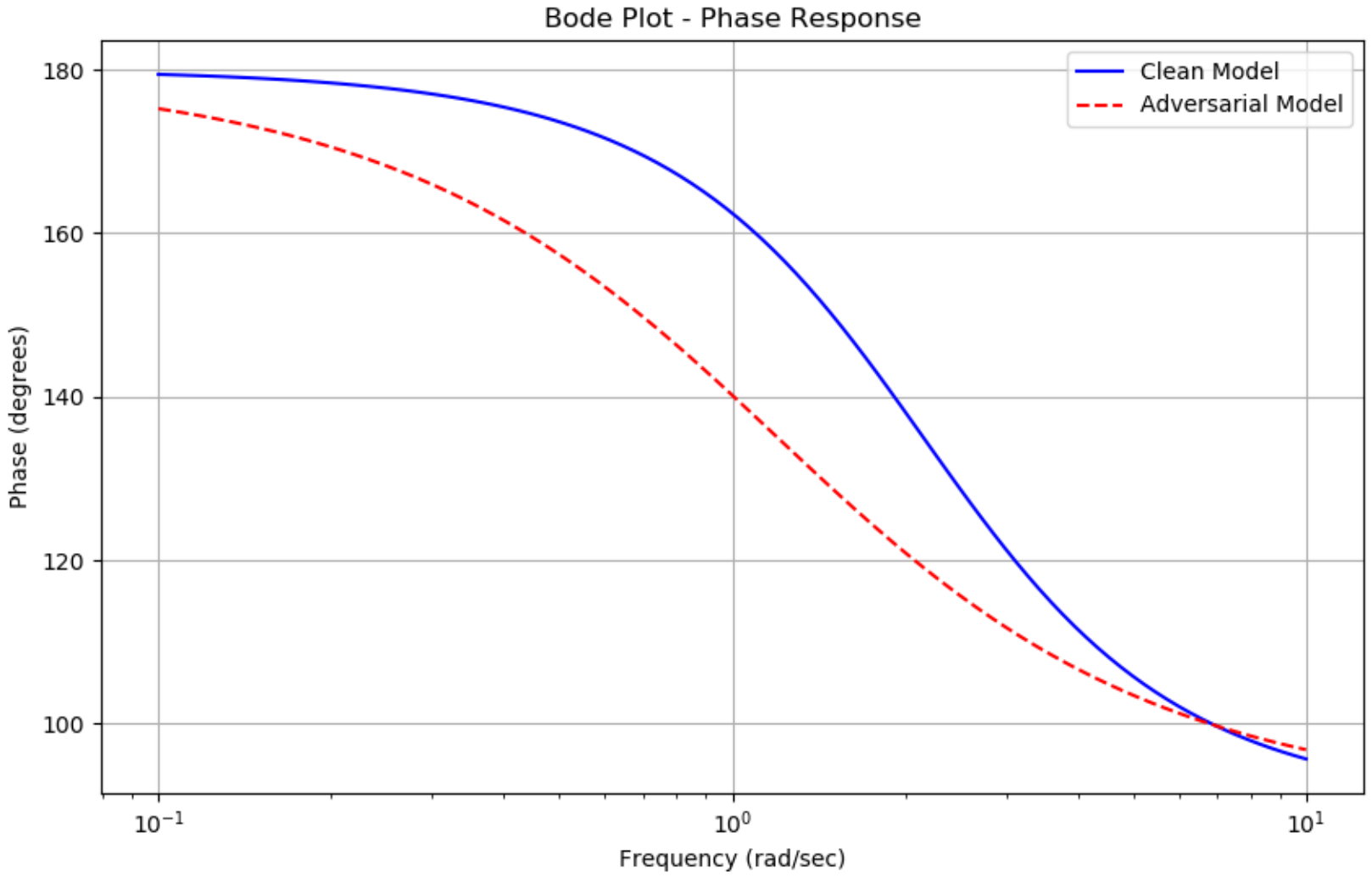


Ref : Improvement of a Parallel Type Two-axial Actuator Controlled by a Multi-layered Neural Network

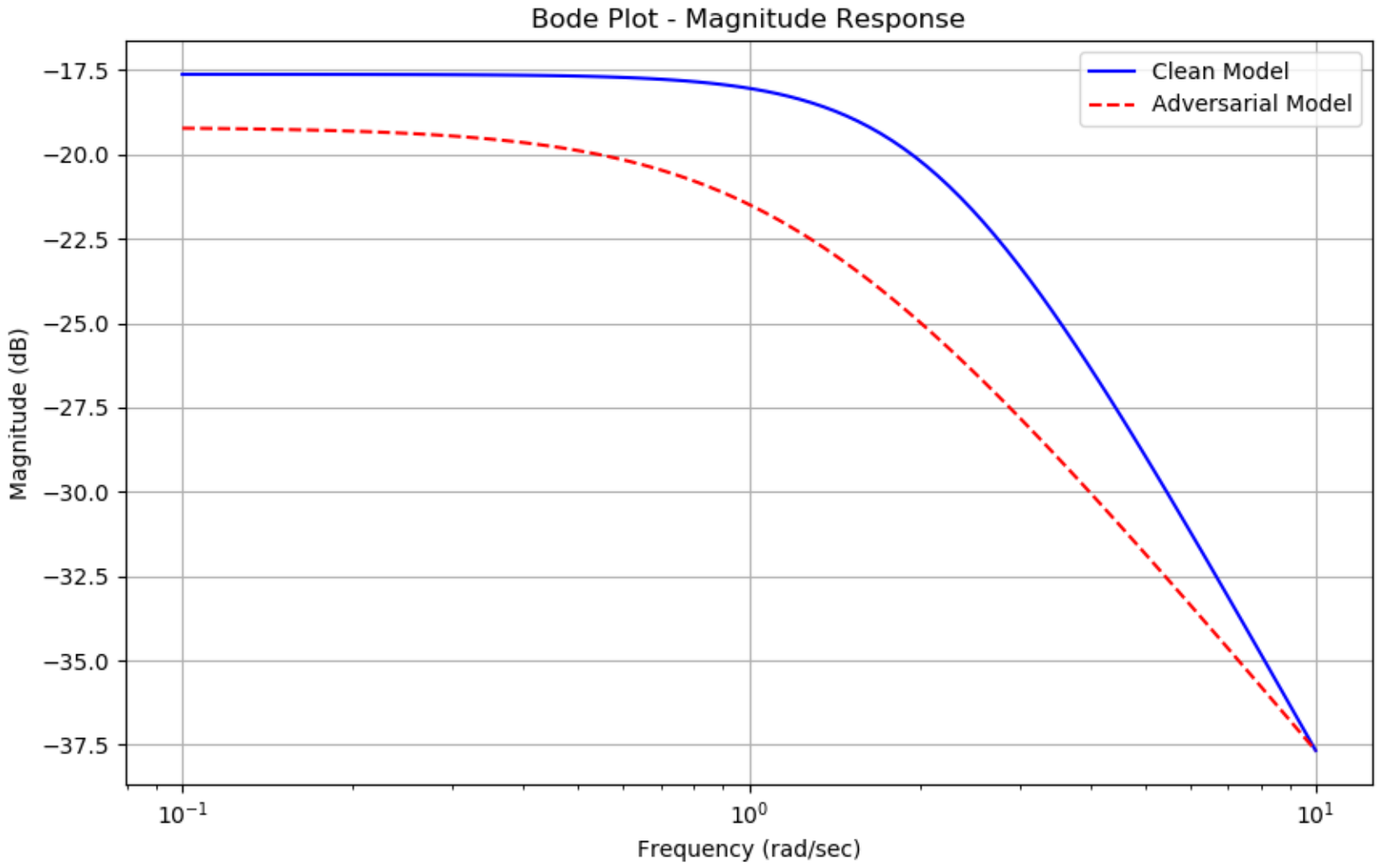
Stay Negative



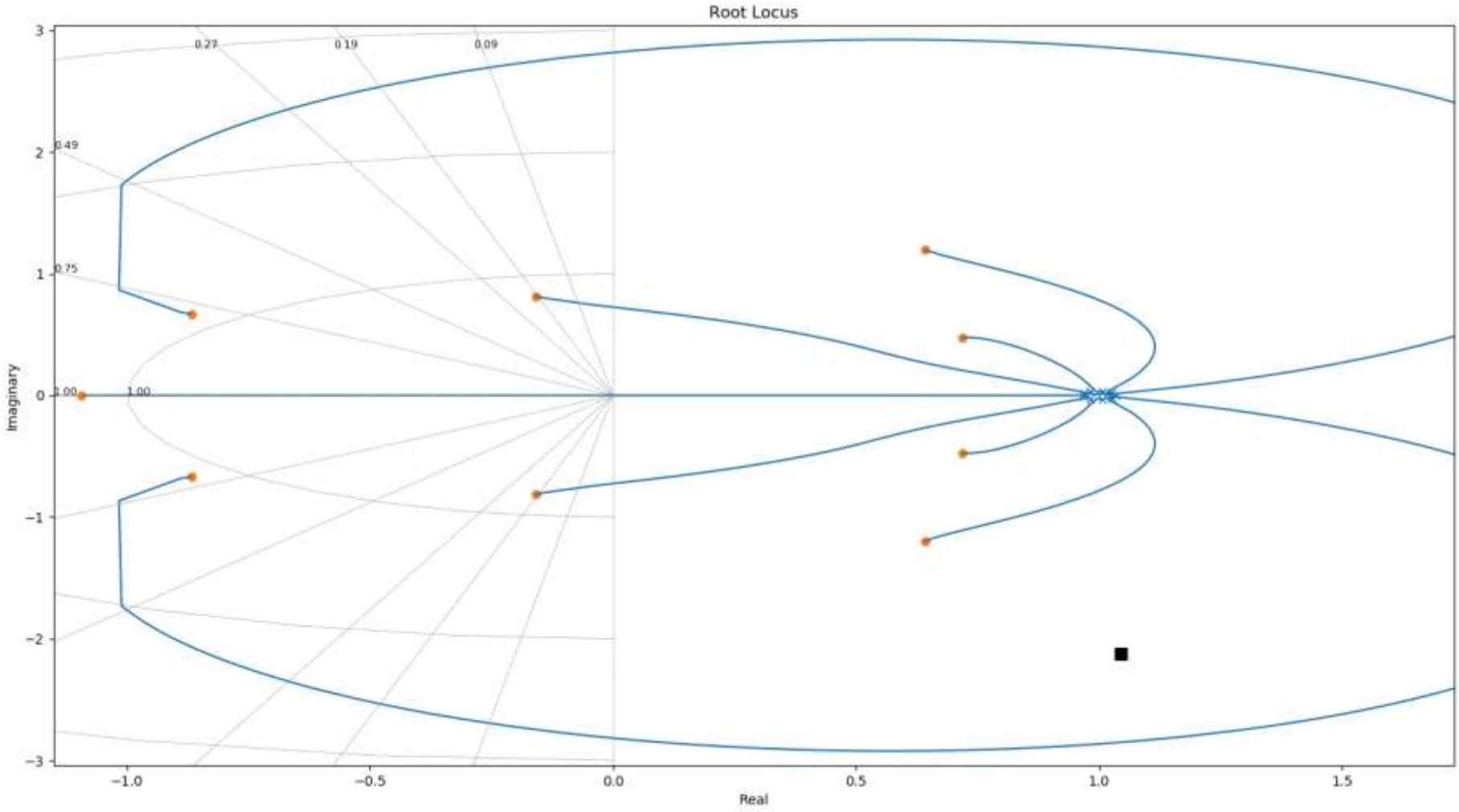
Chase The Race



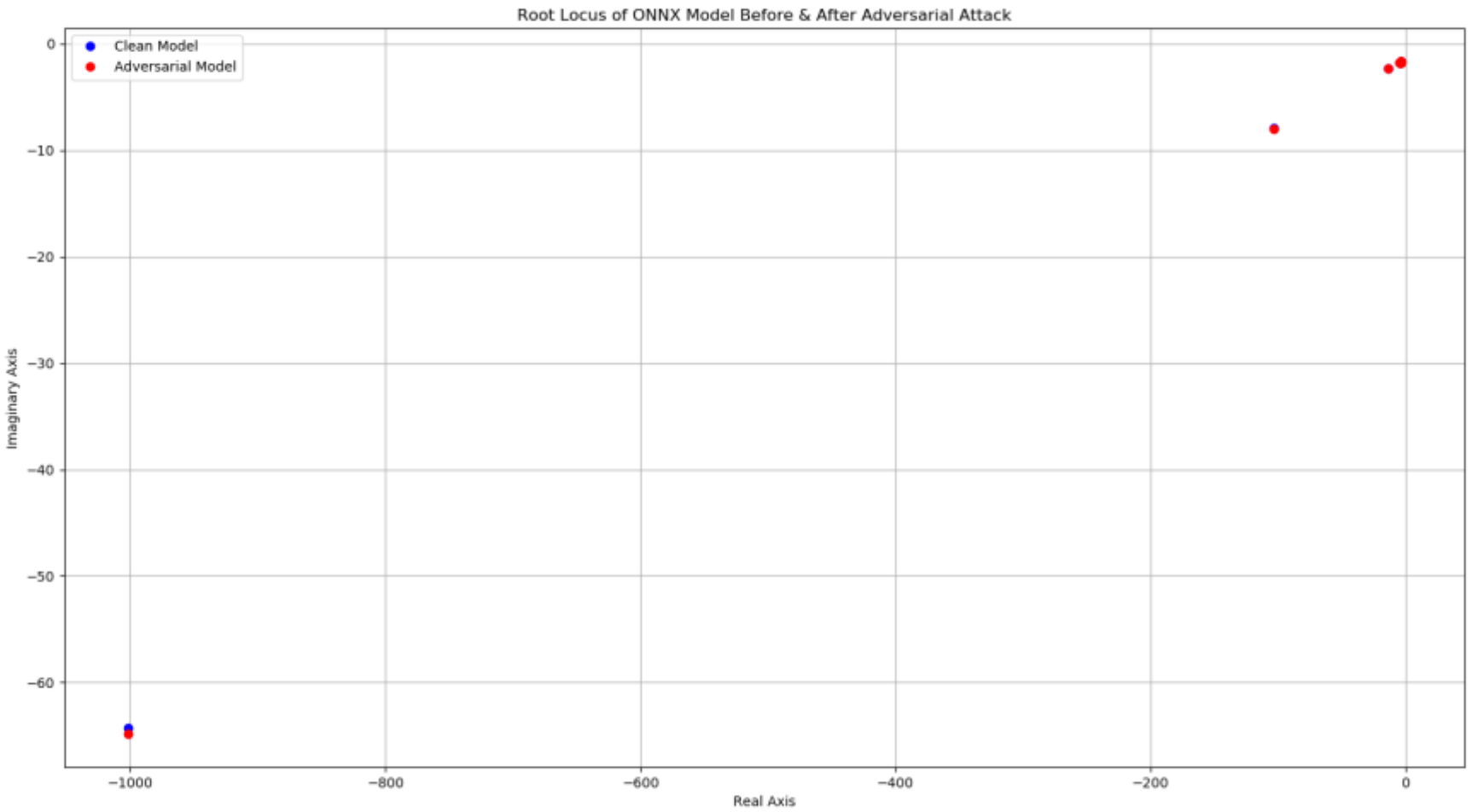
Fall Off


Bode Plot - Magnitude Response
Clean Model
Adversarial Model
−17.5
−20.0
−22.5
−25.0
−27.5
−30.0
−32.5
−35.0
−37.5
Magnitude (dB)
10^{-1}
10^{0}
10^{1}
Frequency (rad/sec)

Whole Body Vibration Exercise



Final Attack



Contribution

- Tenzzer – Fuzzing Safetensors
  - https://github.com/Yashodhanvivek/Tenzzer
- Torcher – Fuzzing and poisoning PyTorch Model

https://github.com/Yashodhanvivek/Torcher

Demo