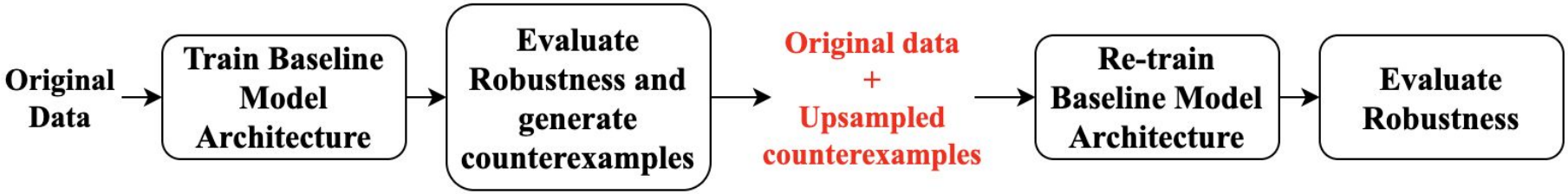
Improving Robustness based on Marabou counterexamples

Vaidehi Joshi
Carnegie Mellon University

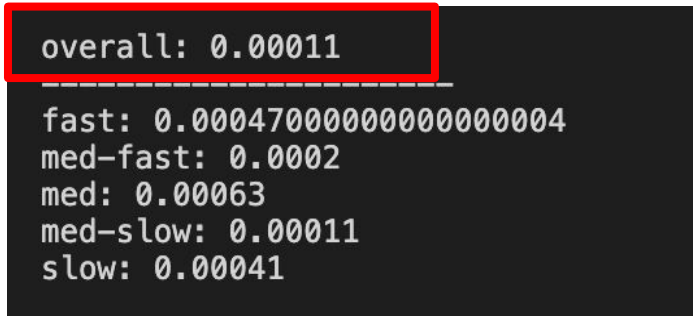
Motivation



Baseline model:
Epsilon: 0.00011

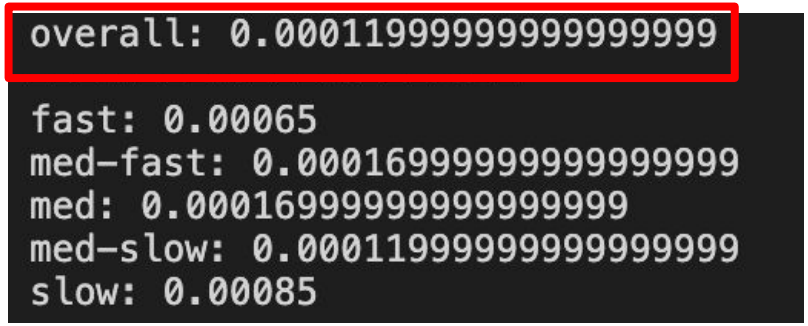
Upsampling technique 1

Robustness counterexamples



Original data >>> upsampled data

Robustness + clustering counterexamples



Original data : upsampled data = 1:1

- Upsampling: Choose a **random number** between **[ε , ε + 0.01]** based on epsilon of each counter example. Further, **add** and **subtract** that random number to the counterexample.
- Generate *n* examples for one counterexample.
- Finally, re-train model with **Original data + Upsampled data**

Statistics of verified epsilons for Upsampling technique 1

Baseline model →

	count	mean	min	5%	25%	50%	75%	95%	delta max
fast	517.0	0.028733	0.00030	0.002552	0.01064	0.02414	0.03599	0.062524	0.26471
med-fast	473.0	0.026576	0.00015	0.003330	0.01116	0.01580	0.02184	0.104452	0.16012
med	493.0	0.027524	0.00011	0.002260	0.00885	0.01996	0.04116	0.070700	0.13579
med-slow	521.0	0.023497	0.00040	0.003090	0.00870	0.01856	0.02931	0.046190	0.65801
slow	479.0	0.026293	0.00033	0.003520	0.01285	0.02403	0.03403	0.052830	0.19719

Re-trained model on original + upsampled data →

spred	count	mean	std	min	25%	50%	75%	95%	max
0	500.0	0.028161	0.025328	0.0	0.009598	0.02337	0.041857	0.065970	0.18326
1	500.0	0.021727	0.020214	0.0	0.007543	0.01861	0.030080	0.053873	0.19010
2	500.0	0.015916	0.017521	0.0	0.001605	0.01174	0.023130	0.048970	0.10794
3	500.0	0.021678	0.023616	0.0	0.008558	0.01880	0.029495	0.048905	0.27429
4	500.0	0.019120	0.016736	0.0	0.008568	0.01680	0.024530	0.051026	0.18059

Upsampling technique 2

Robustness counterexamples

```
overall: 0.00011
----------------------
fast: 0.0003
med-fast: 0.00011999999999999999
med: 0.00011
med-slow: 0.00025
slow: 0.00021
```

Original data >>> upsampled data

Robustness + clustering counterexamples

```
overall: 0.00011999999999999999
----------------------
fast: 0.0002
med-fast: 0.00042
med: 0.00011999999999999999
med-slow: 0.0002
slow: 0.0005
```

Original data : upsampled data = 2:1

- Upsampling: Duplicate every counterexample *n* times
- Re-train model with **Original data + Upsampled data**

Statistics of verified epsilons for Upsampling technique 2

Baseline model →

	count	mean	min	5%	25%	50%	75%	95%	delta max
fast	517.0	0.028733	0.00030	0.002552	0.01064	0.02414	0.03599	0.062524	0.26471
med-fast	473.0	0.026576	0.00015	0.003330	0.01116	0.01580	0.02184	0.104452	0.16012
med	493.0	0.027524	0.00011	0.002260	0.00885	0.01996	0.04116	0.070700	0.13579
med-slow	521.0	0.023497	0.00040	0.003090	0.00870	0.01856	0.02931	0.046190	0.65801
slow	479.0	0.026293	0.00033	0.003520	0.01285	0.02403	0.03403	0.052830	0.19719

Re-trained model on original + upsampled data →

spred	count	mean	std	min	25%	50%	75%	95%	max
0	500.0	0.028161	0.025328	0.0	0.009598	0.02337	0.041857	0.065970	0.18326
1	500.0	0.021727	0.020214	0.0	0.007543	0.01861	0.030080	0.053873	0.19010
2	500.0	0.015916	0.017521	0.0	0.001605	0.01174	0.023130	0.048970	0.10794
3	500.0	0.021678	0.023616	0.0	0.008558	0.01880	0.029495	0.048905	0.27429
4	500.0	0.019120	0.016736	0.0	0.008568	0.01680	0.024530	0.051026	0.18059

Future work

- Explore into different ways to create adversarial samples using the counterexamples
- Try: reinforcement technique (assigning weight to every counterexample)