ML-LOO: Detecting Adversarial Examples with Feature Attribution

Puyudi Yang∗ , Jianbo Chen† , Cho-Jui Hsieh‡ , Jane-Ling Wang∗ , Michael I. Jordan†

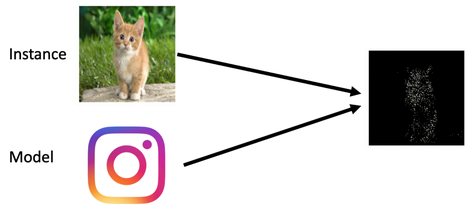
University of California, Davis∗, University of California, Berkeley†, University of California, Los Angeles‡

ABSTRACT

Deep neural networks obtain state-of-the-art performance on a series of tasks. However, they are easily fooled by adding a small adversarial perturbation to the input. The perturbation is often imperceptible to humans on image data. We observe a significant difference in feature attributions between adversarially crafted examples and original examples. Based on this observation, we introduce a new framework to detect adversarial examples through thresholding a scale estimate of feature attribution scores. Furthermore, we extend our method to include multi-layer feature attributions in order to tackle attacks that have mixed confidence levels. As demonstrated in extensive experiments, our method achieves superior performances in distinguishing adversarial examples from popular attack methods on a variety of real data sets compared to state-of-the-art detection methods. In particular, our method is able to detect adversarial examples of mixed confidence levels, and transfer between different attacking methods. We also show that our method achieves competitive performance even when the attacker has complete access to the detector.

FEATURE ATTRIBUTIONS

For a given instance, assign a vector of importance scores for each feature.



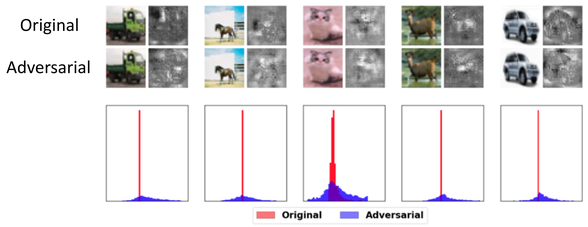
- Model

$$f : \mathbb{R}^d \to [0,1]^C$$

- Leave One Out (LOO) feature attributions (Li, Monroe, and Jurafsky 2016)

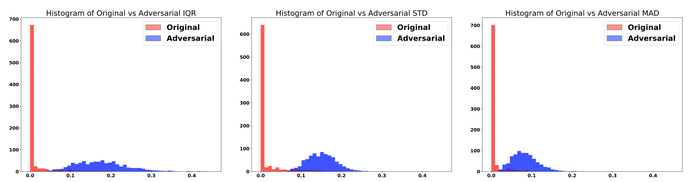
$$\phi(x)_i := f(x)_c - f(x_{(i)})_c, \text{ where } c = \arg\max_{j \in C} f(x)_j$$

A DIFFERENCE IN ATTRIBUTION MAPS

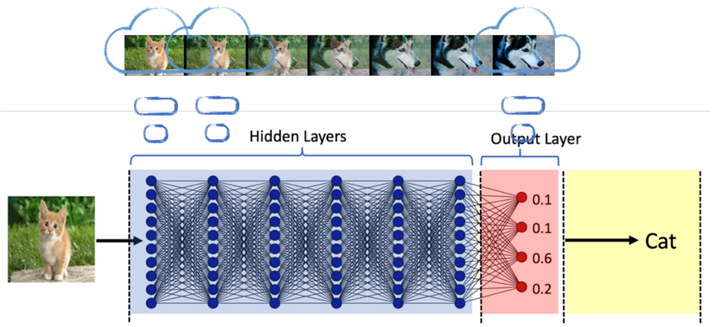


HISTOGRAM OF LOO DISPERSION

- STD (standard deviation)
- MAD (median absolute deviation)
  - median of absolute differences between entries and their median
- **IQR** (Interquartile range)
  - difference between the 75th percentile and the 25th percentile



EXTENSION TO MULTI-LAYER LOO



Procedure:

Step 1: Compute IQR of LOO scores for each layer

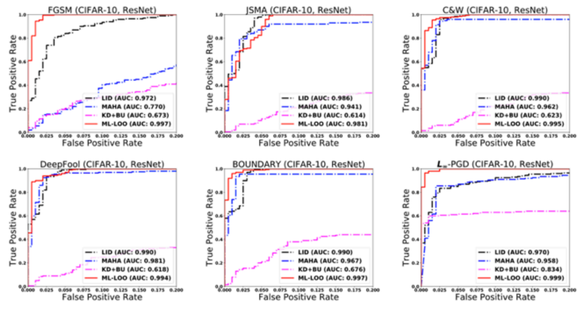
Step 2: Fit a logistic regression

CONCLUSION

- A new framework: Detecting adversarial examples with multi-layer feature attribution.
- State-of-the-art performance: Outperforming other methods in detecting various kinds of attacks, including varied confidence levels and transfer attack.
- White-box threat model: Achieving competitive performance (See our paper for details).

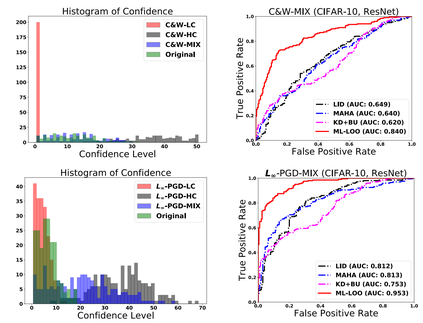
PERFORMANCE ON KNOWN ATTACKS

- FGSM (Goodfellow, Shlens, and Szegedy 2014).
- L^∞- PGD (Kurakin, Goodfellow and Bengio 2017; Madry et al. 2018).
- C&W (Carlini and Wagner 2017).
- Deep-Fool (Moosavi-Dezfooli, Fawzi, and Frossard 2016).
- Boundary Attack (Brendel, Rauber, and Bethge 2018).
- JSMA (Papernot et al. 2016).



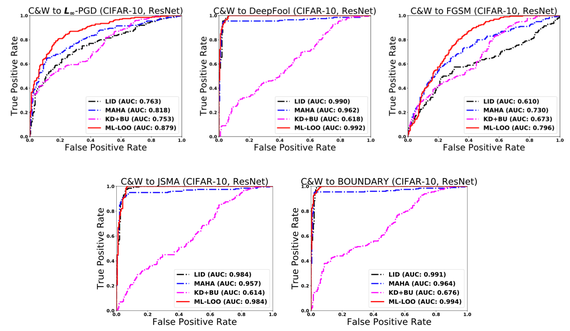
PERFORMANCE ON MIXED-CONFIDENCE ATTACKS

- Evaluate detectors on adversarial examples with mixed confidence levels (Lu, Chen, Yu 2018; Athalye, Carlini, Wagnar 2018).



PERFORMANCE ON TRANSFERRED ATTACKS

- Train ML-LOO on CW attack and evaluate it on other attacks.



Berkeley