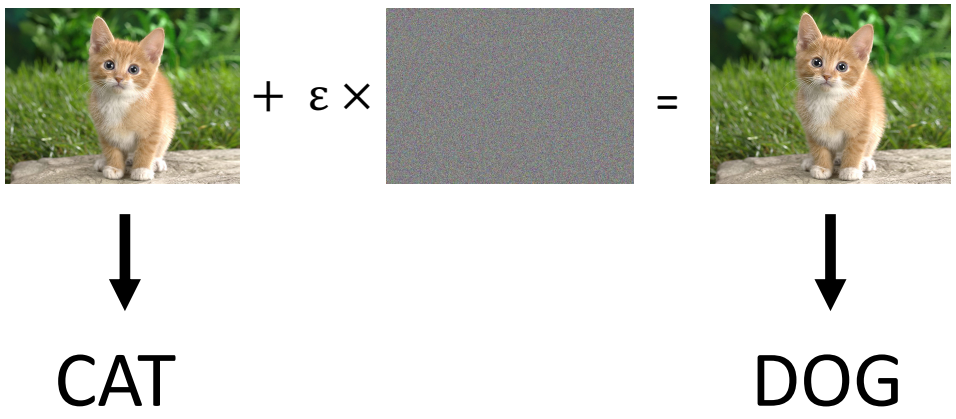
ML-LOO: Detecting Adversarial Examples with Feature Attribution

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

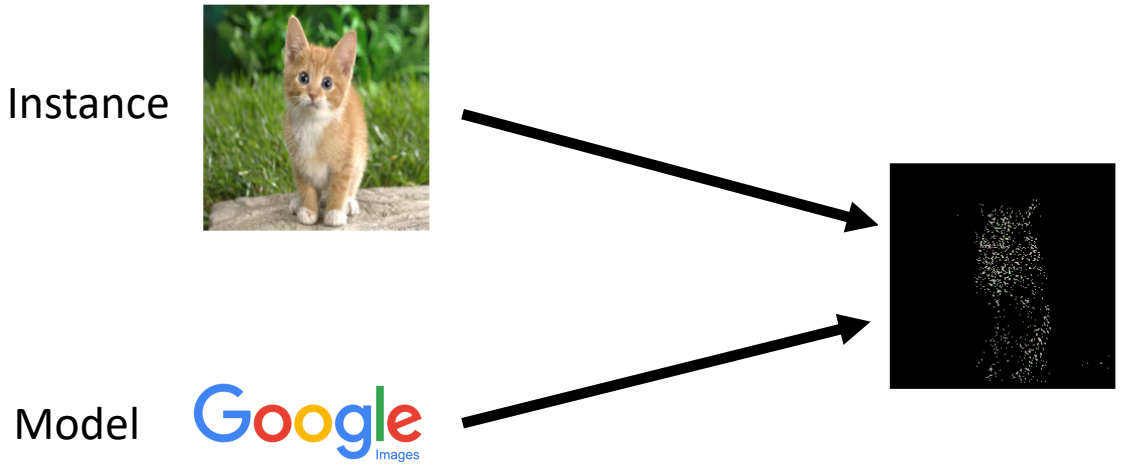
University of California, Davis*, University of California, Berkeley†, University of California, Los Angeles‡

Adversarial attack (Szegedy et. al. 2013)

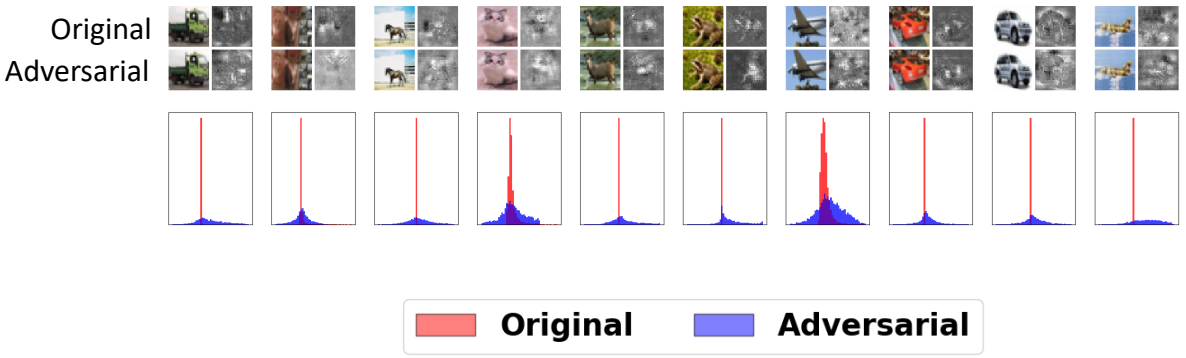
Fool a (deep) image classifier with minimum perturbation.



Feature attribution

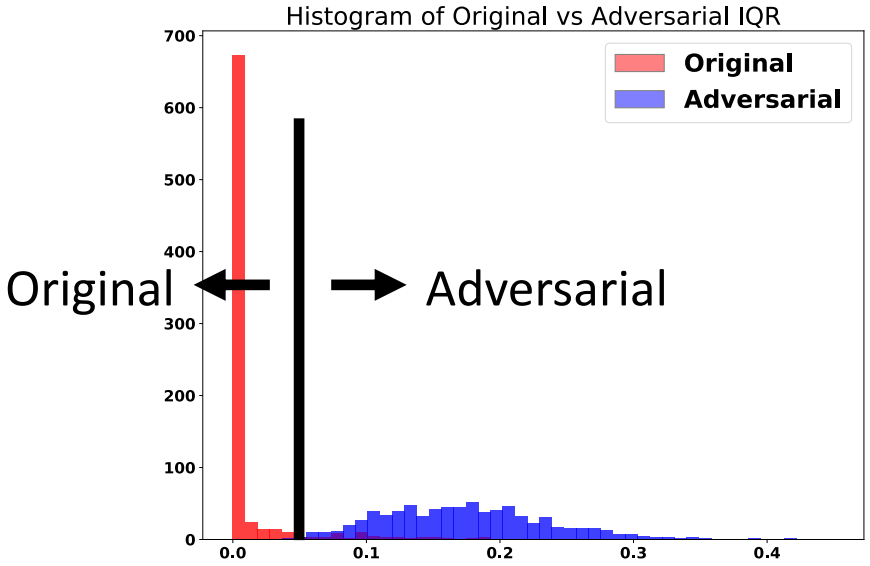


A difference in attribution maps



Histogram of Attribution Scores

A universal scaling difference



Histogram of Attribution Interquartile Range