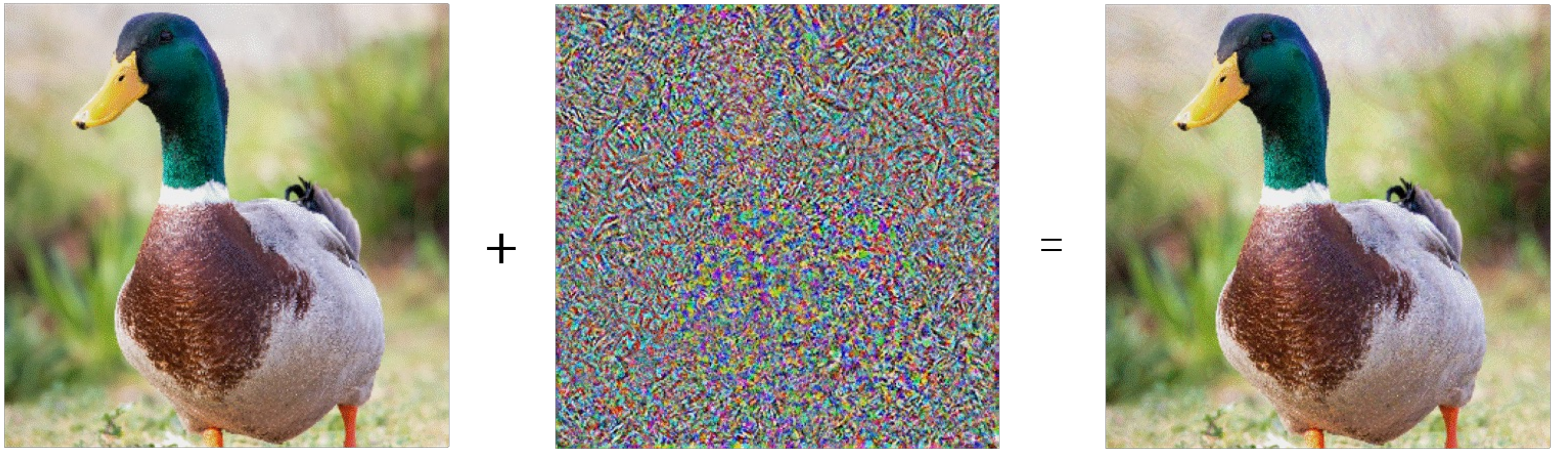
Adversarial

Eric Wong

9/8/2022

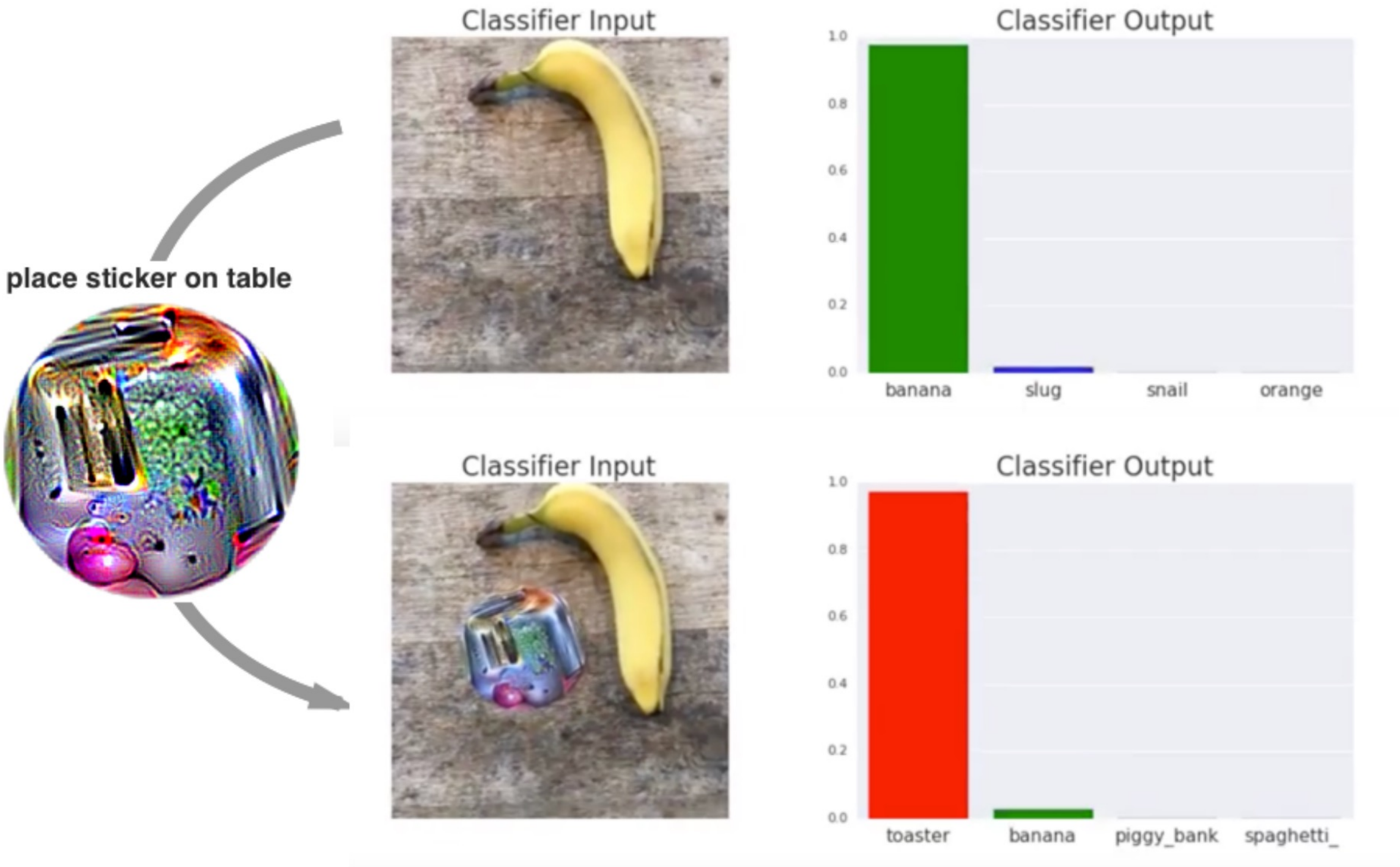
Noise attack



“Duck”

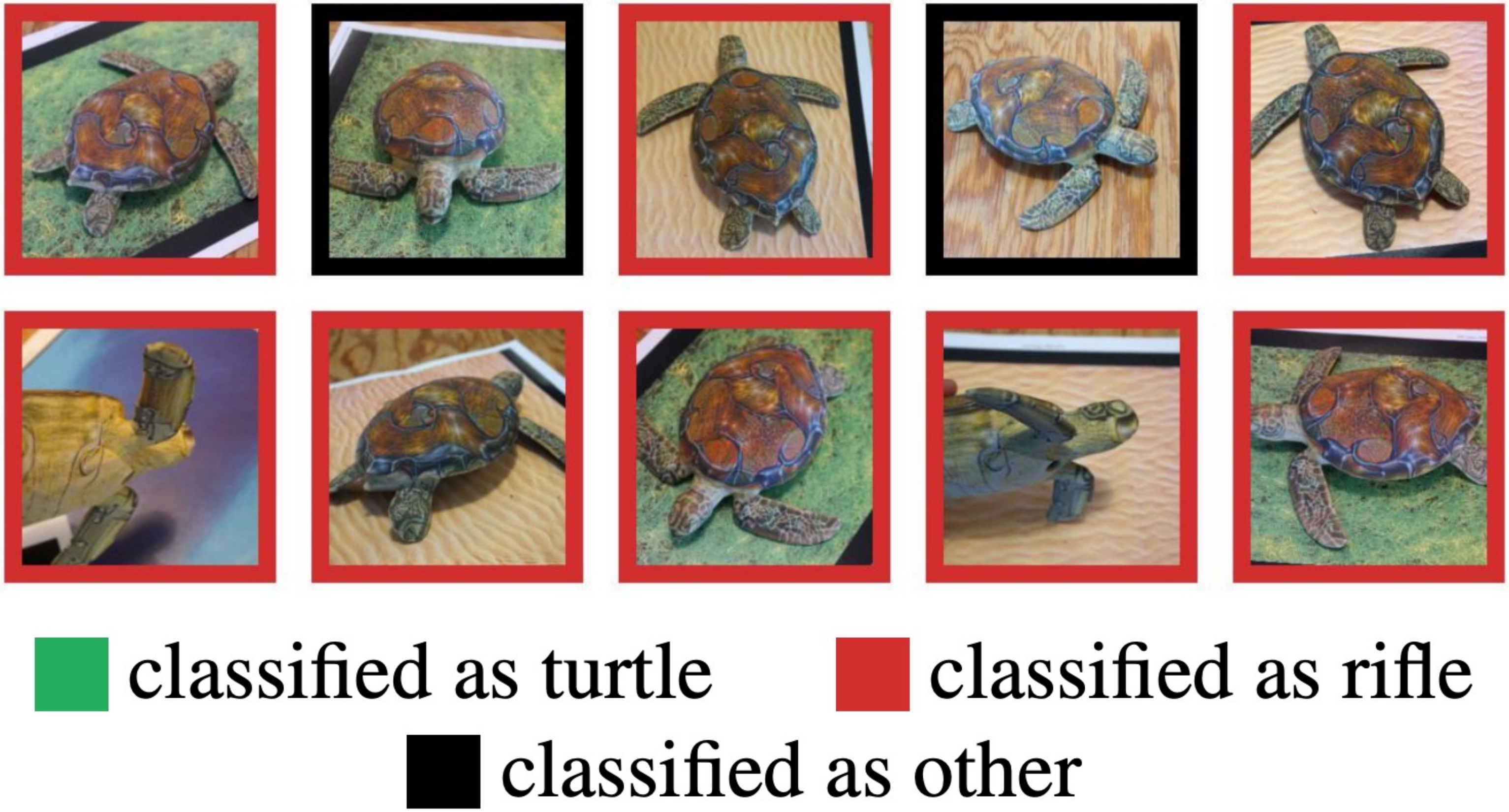
“Hermit crab”

Patch attack



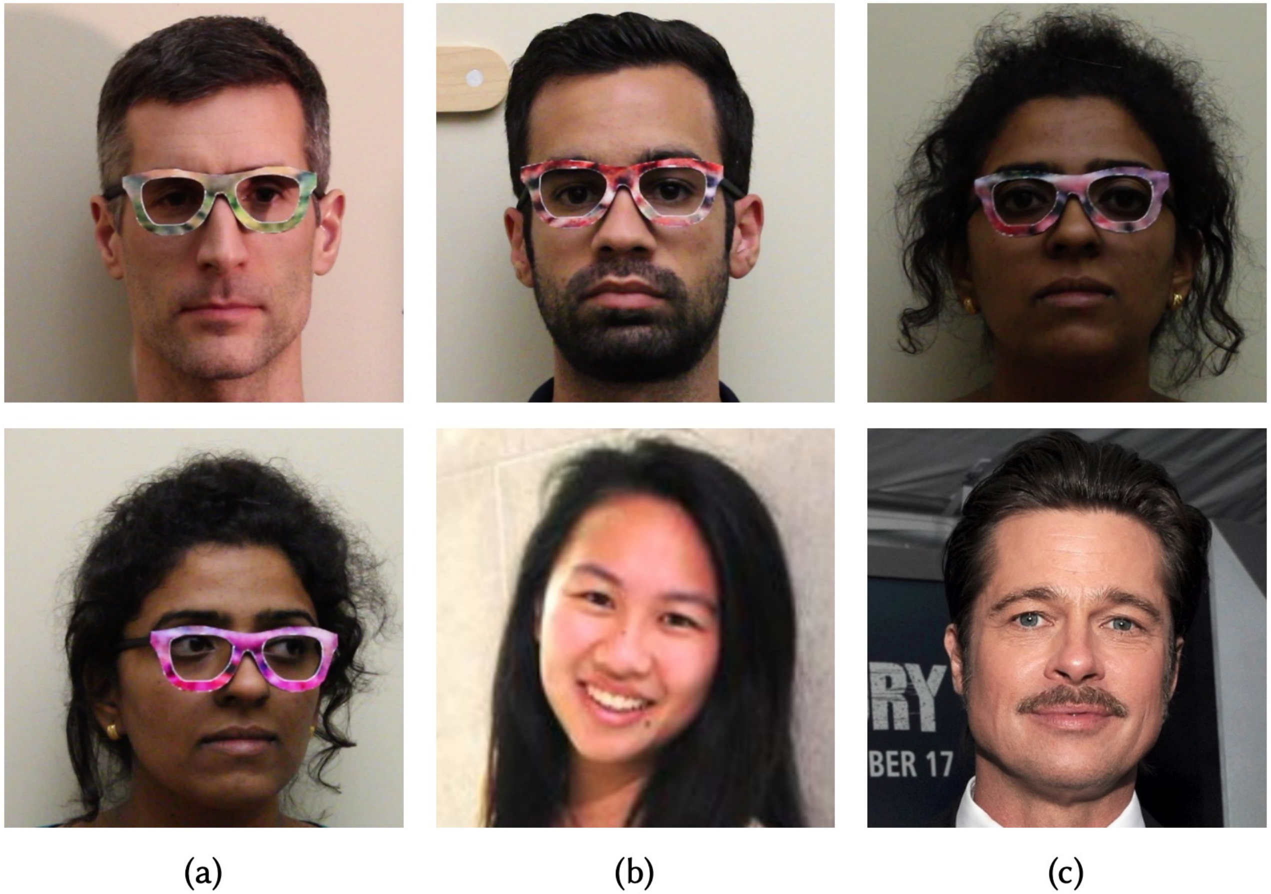
Brown et al. 2017 “Adversarial Patch”

3D printed textures



Athalye et al. 2017 “Synthesizing Robust Adversarial Examples”

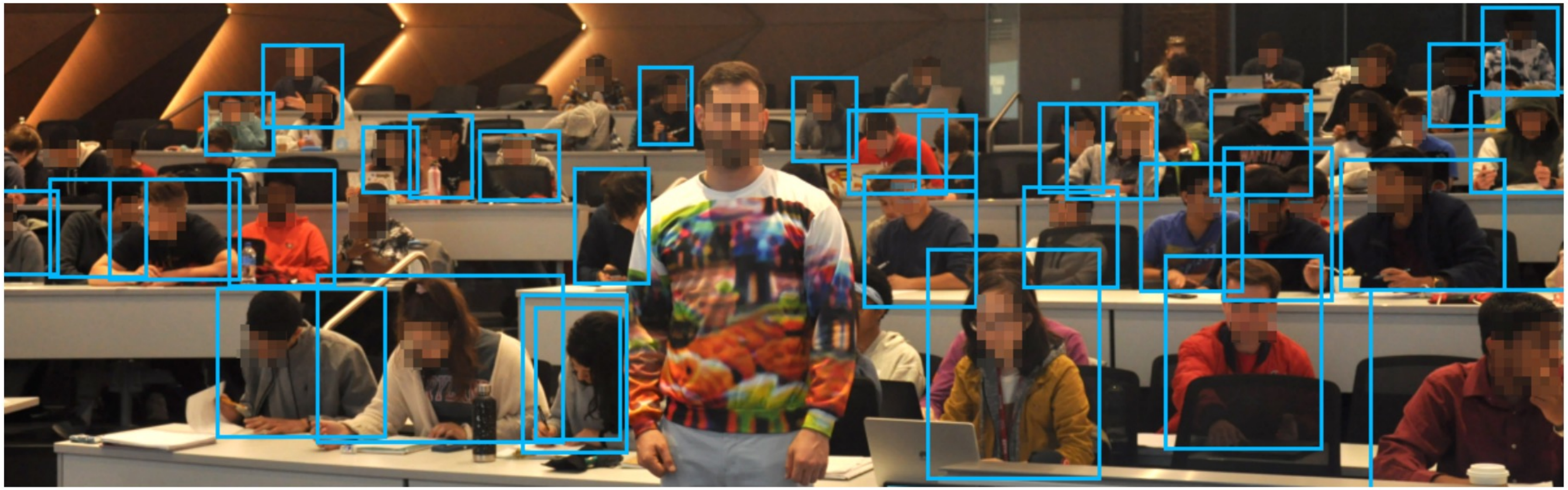
Glasses



(a) (b) (c)

Sharif et al. 2018 “A General Framework for Adversarial Examples with Objectives”

Clothing

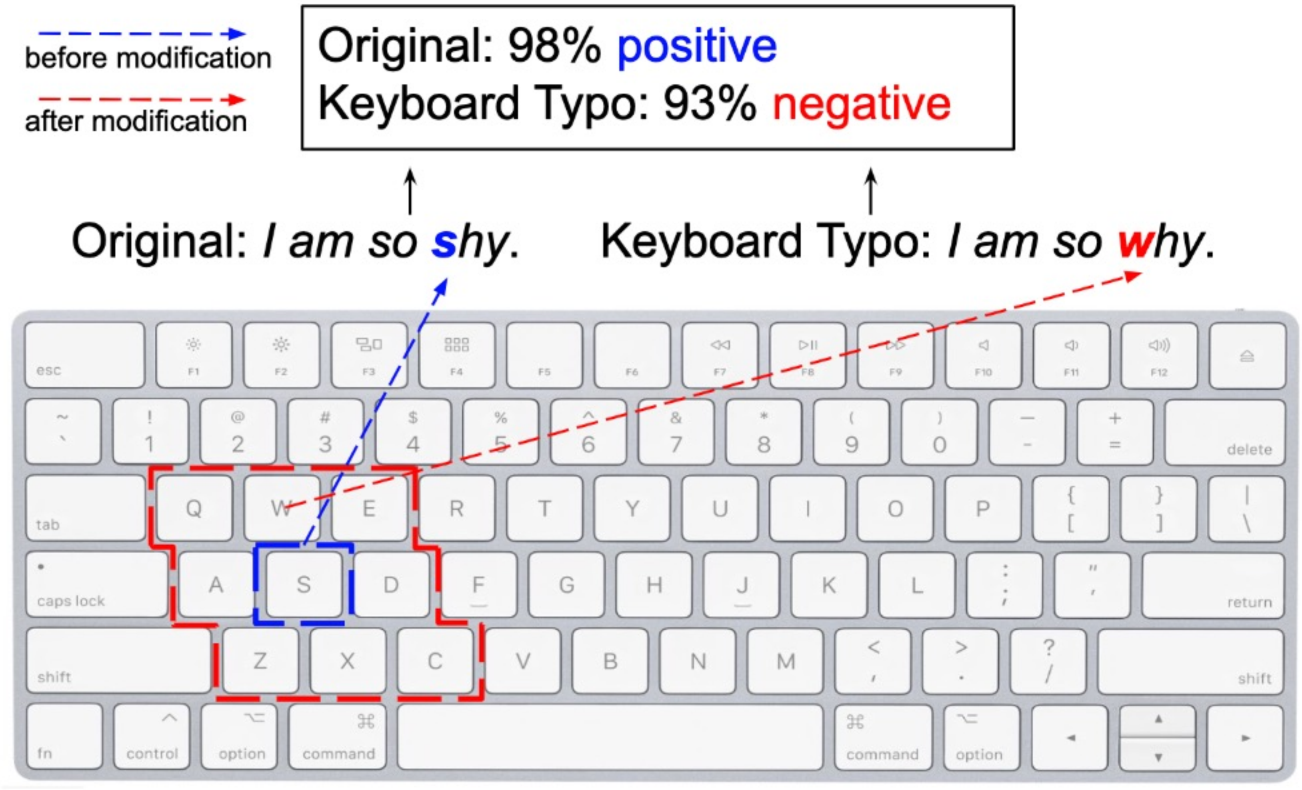


Wu et al. 2019 “Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors”

Camera stickers

Li et al. 2019 “Adversarial camera stickers: A physical camera based attack on deep learning sytems”

Typos

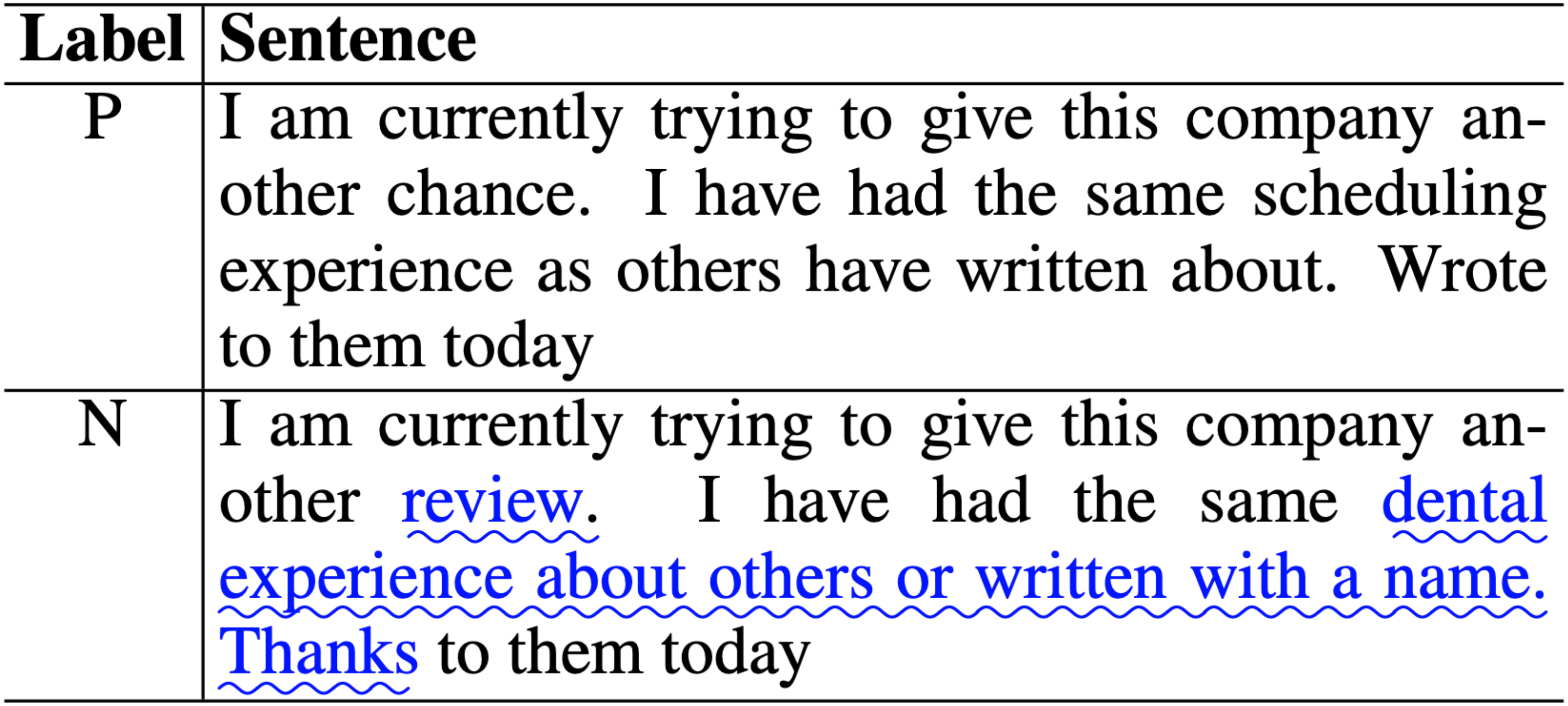

before modification
after modification
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12
~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 — - + = delete
tab Q W E R T Y U I O P { [}] | \
caps lock A S D F G H J K L : ; " ' return
shift Z X C V B N M < , > . ? / shift
fn control option command command option

Original: 98% positive
Keyboard Typo: 93% negative

Original: *I am so **shy**.*
Keyboard Typo: *I am so **why**.*

Sun et al. 2020 “Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT”

Sentences



Label	Sentence
P	I am currently trying to give this company another chance. I have had the same scheduling experience as others have written about. Wrote to them today
N	I am currently trying to give this company another review. I have had the same dental experience about others or written with a name. Thanks to them today

Hsieh et al. 2019 “Natural Adversarial Sentence Generation with Graident based Perturbation”

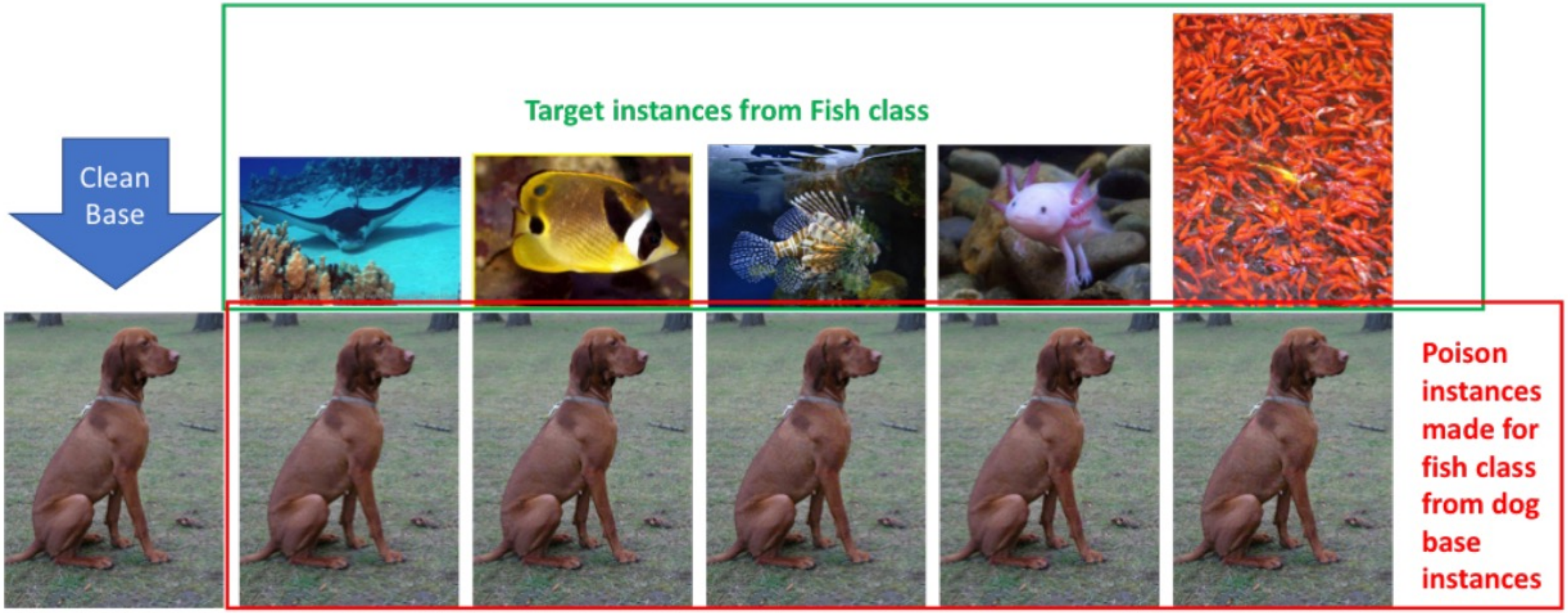
Speech recognition

Abdullah et al. 2020

Data Poisoning

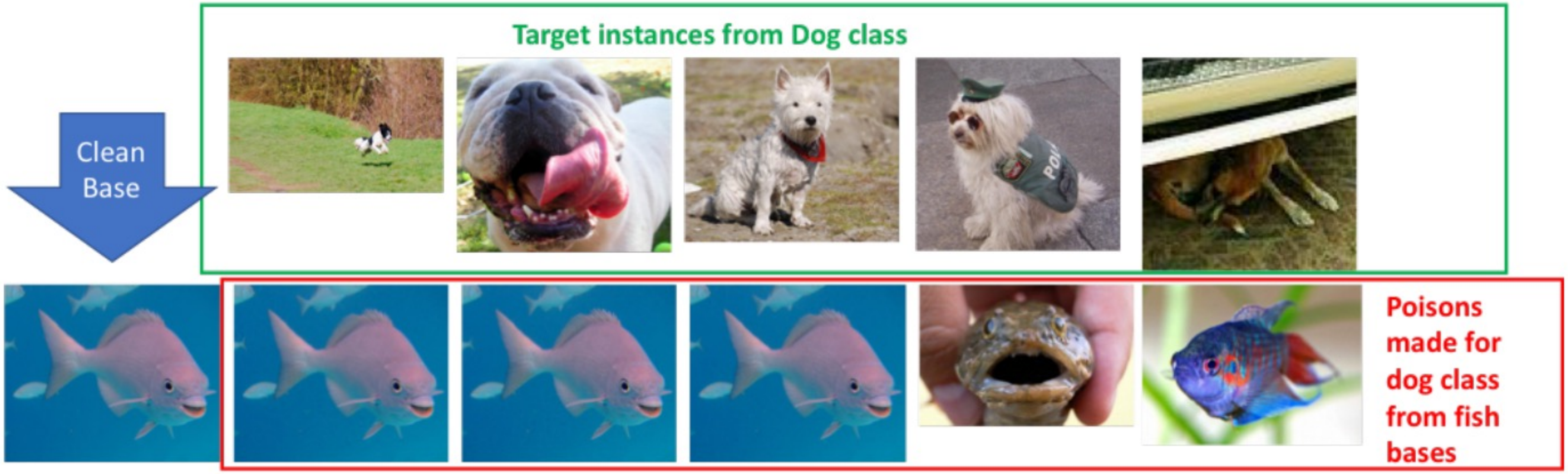
Eric Wong
9/22/2022

“One-shot” poison


Target instances from Fish class
Clean Base
Poison instances made for fish class from dog base instances

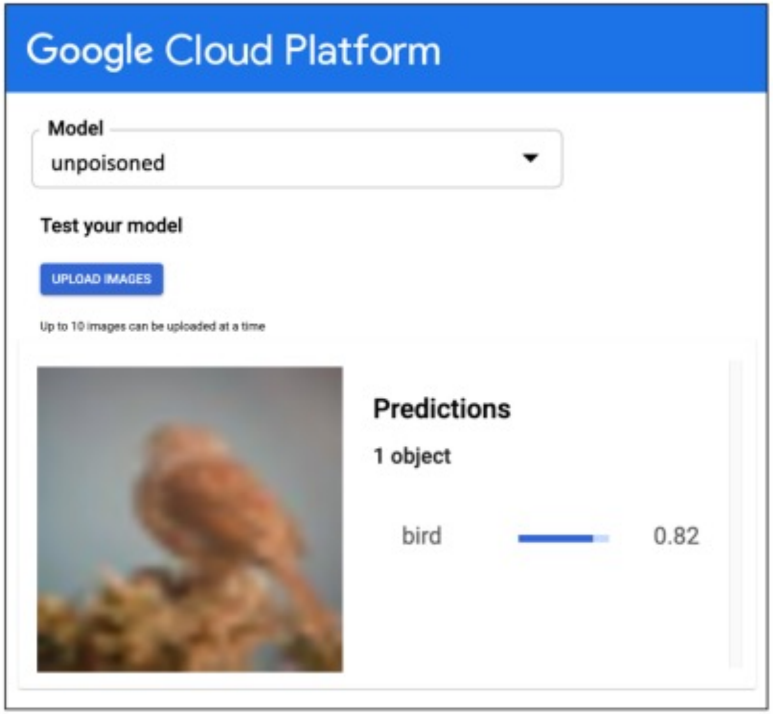
Shafahi et al. 2020 “Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks”

“One-shot” poison



Shafahi et al. 2020 “Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks”

Meta Poisoning



Model
unpoisoned

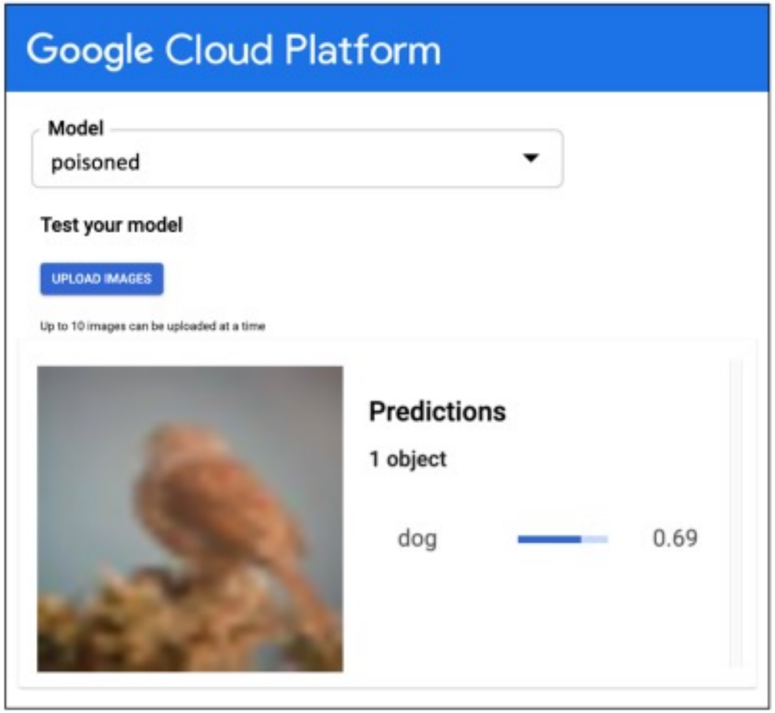
Test your model

Up to 10 images can be uploaded at a time

Predictions

1 object

bird 0.82



Model
poisoned

Test your model

Up to 10 images can be uploaded at a time

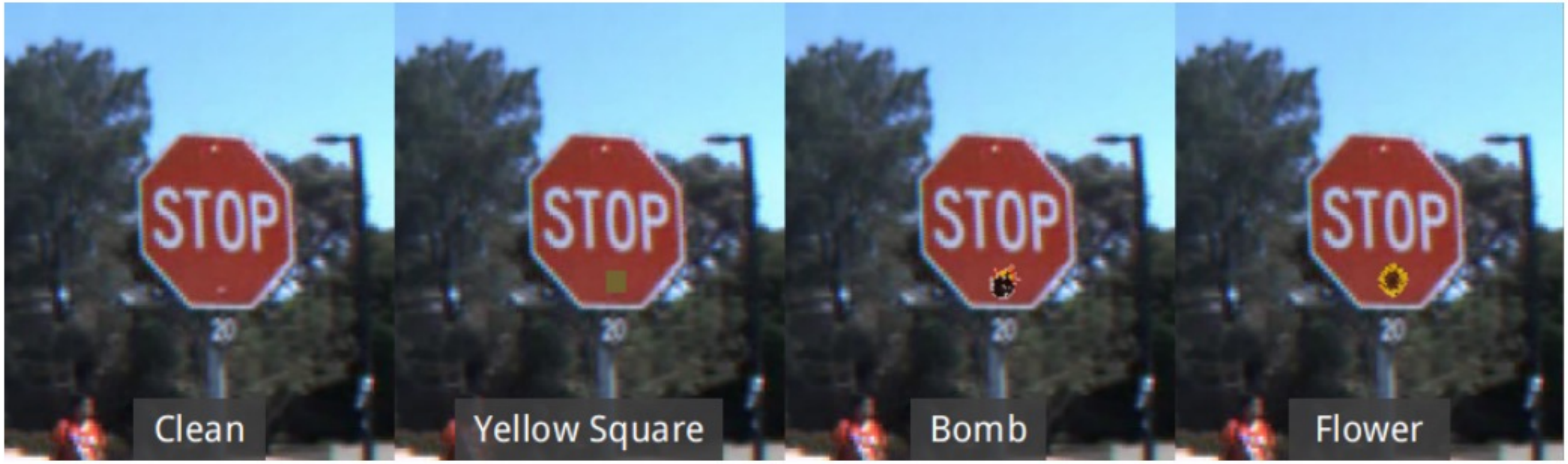
Predictions

1 object

dog 0.69

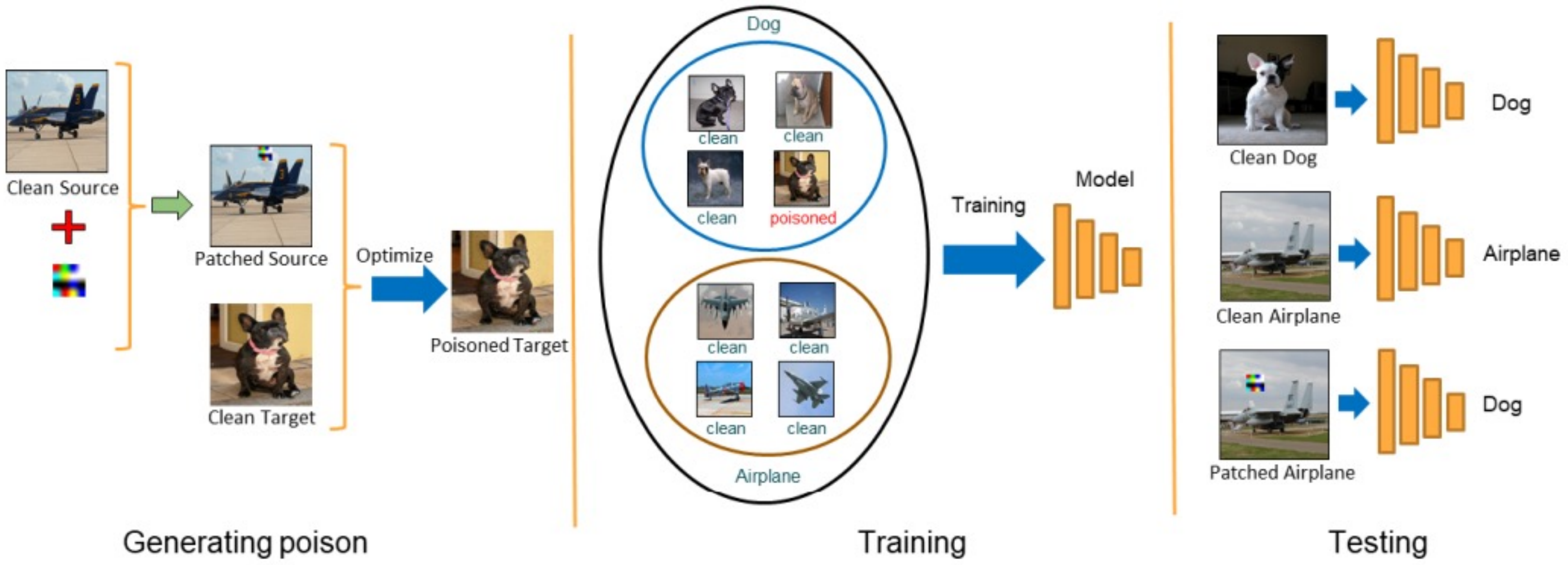
Huang et al. 2020 “MetaPoison: Practical General-purpose Clean-label Data Poisoning”

Backdoor triggers



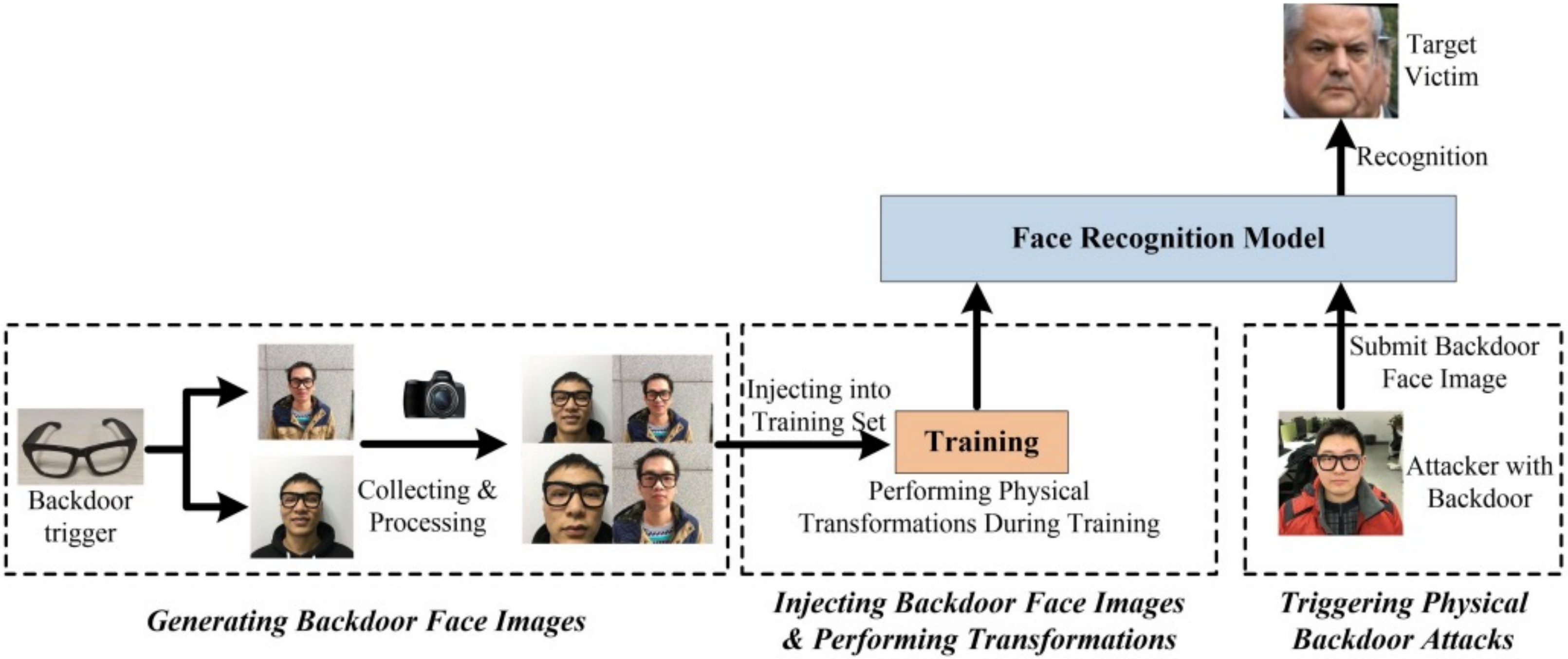
Gu et al. 2017 “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain”

Hidden backdoors



Saha et al. 2019 “Hidden Trigger Backdoor Attacks”

Real & robust backdoors



Distributed backdoor