Testing Robustness of Commercial Online Image Recognition Systems

Riley Blaylock, Lane Colquett, Swapnil Pokhrel

Overview

Three Classifiers

- Google Cloud Vision
- Amazon Rekognition
- Salesforce MetaMind Einstein Vision

Purpose: Generate variety of adversarial images, interact with classifiers’ APIs, record results

Google Cloud Vision - Generating Attacks

Dataset

- 20 predetermined images from imagenet
- Example labels: dog, strawberry, shoe, car

Pretrained Models for generation

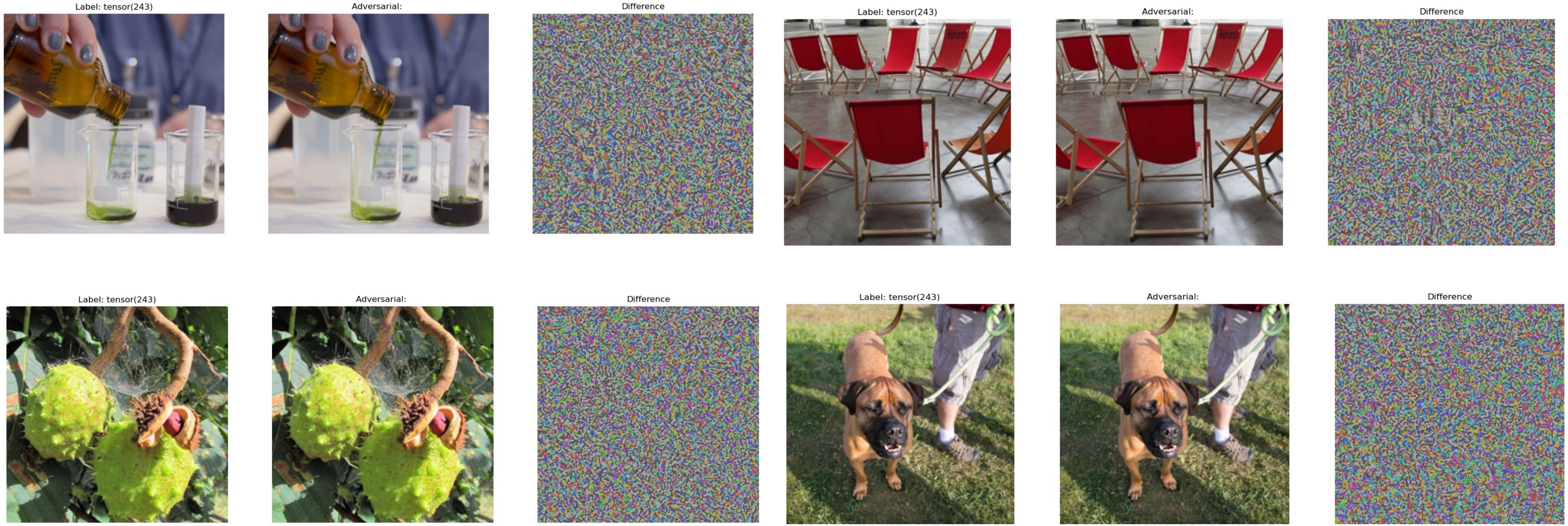
- Resnet34
- GoogLeNet

Attacks - Foolbox

- L-infinity Projected Gradient Descent Attack (PGD)
- Fast Gradient Sign Method (FGSM)
- L-infinity Basic Iterative Method (BIA)

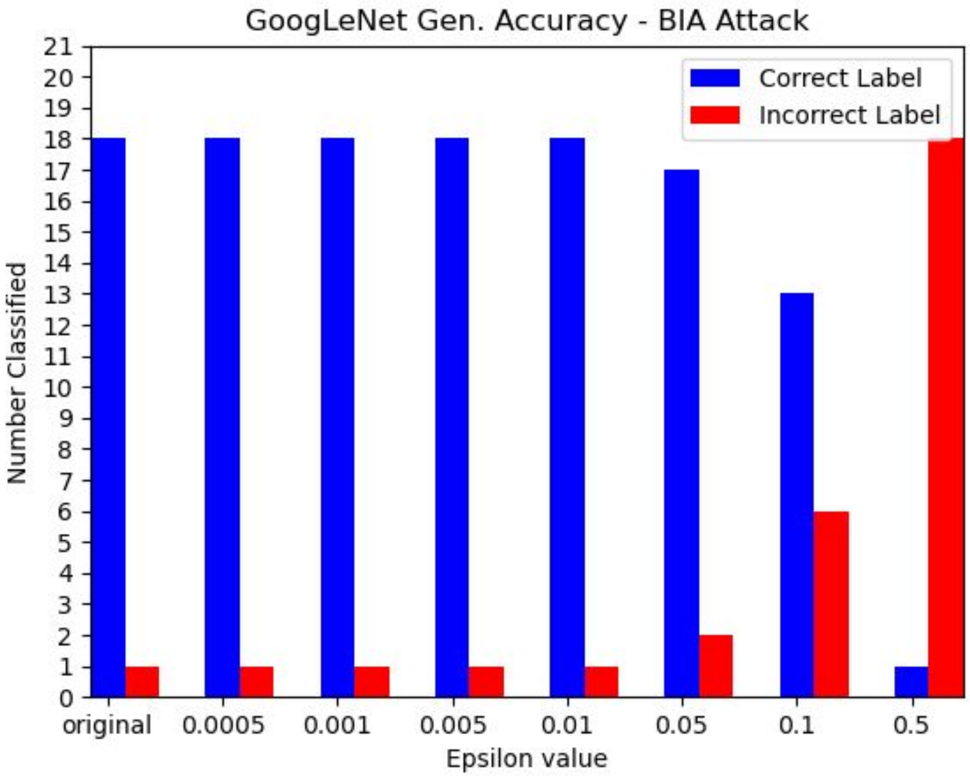
Perturbed Images

Images below: GoogLeNet, PGD attack, Epsilon 0.001

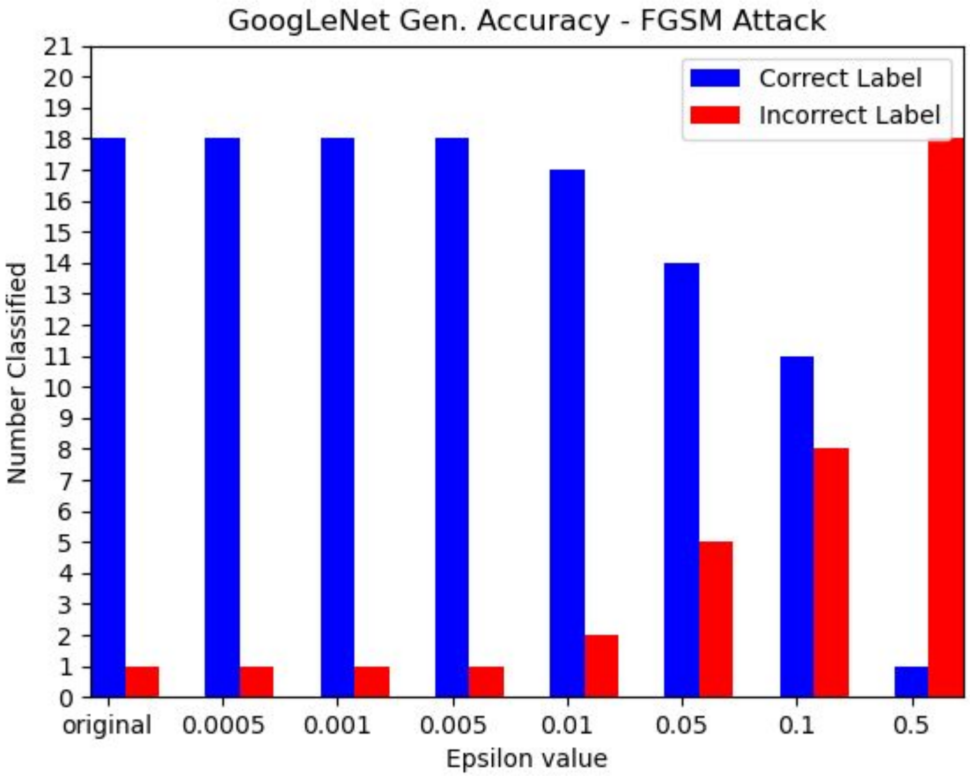


840 total -> 2 pretrained models, 20 base images, 3 generative attacks, 7 perturbation levels -- all generated images result in ~0% accuracy when tested on pretrained model with which they were generated

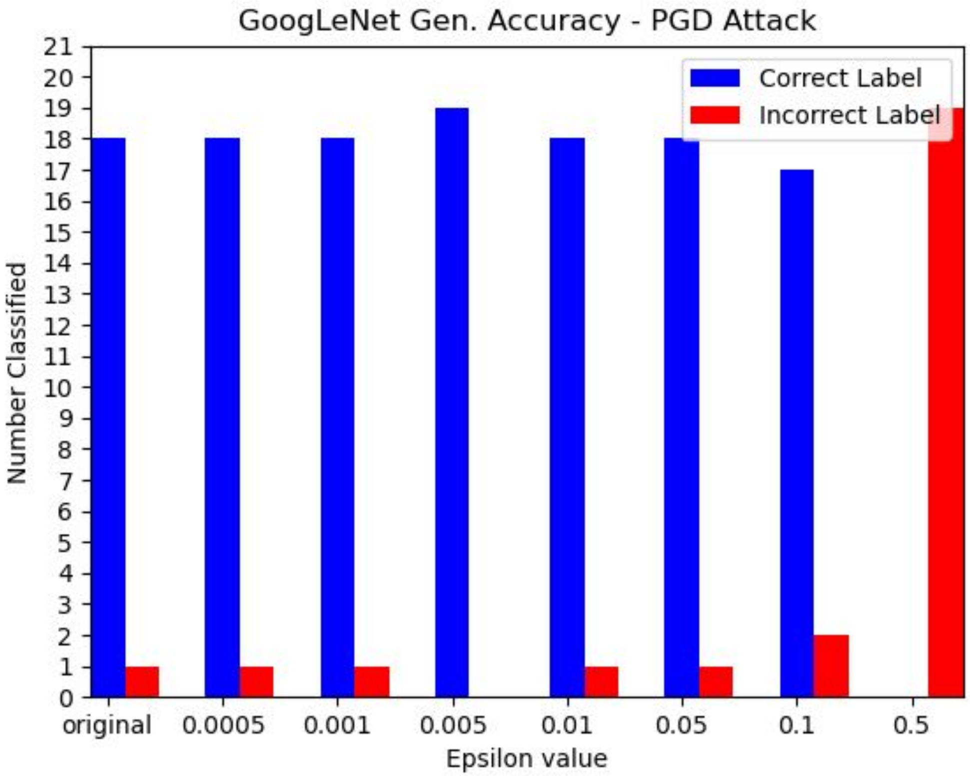
Results - GoogLeNet



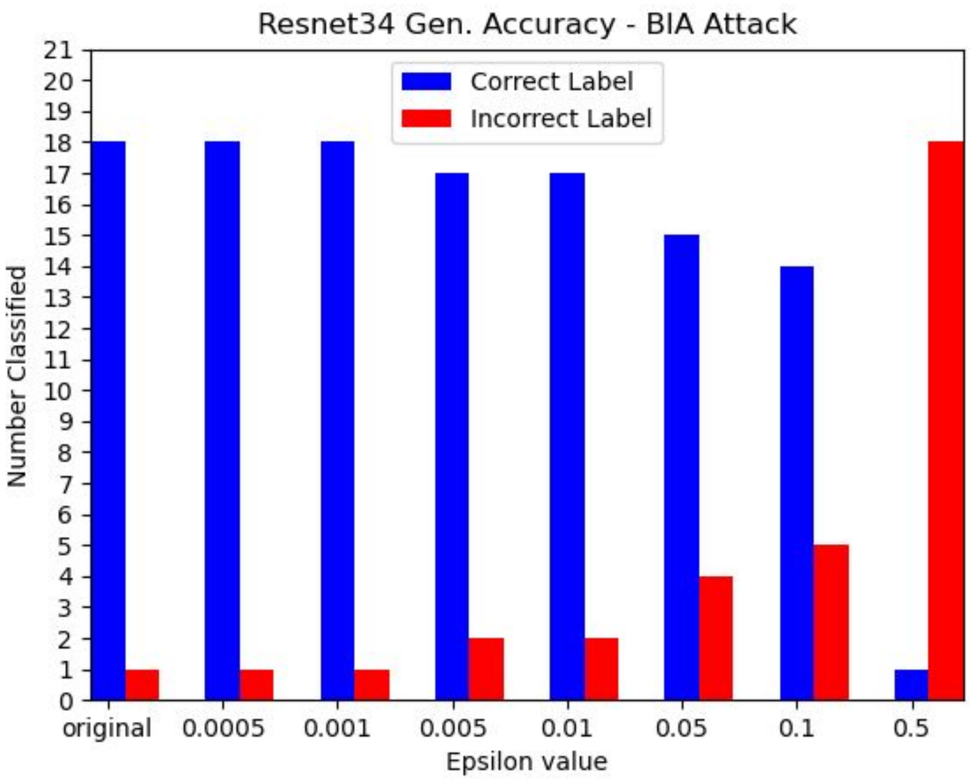
Results - GoogLeNet



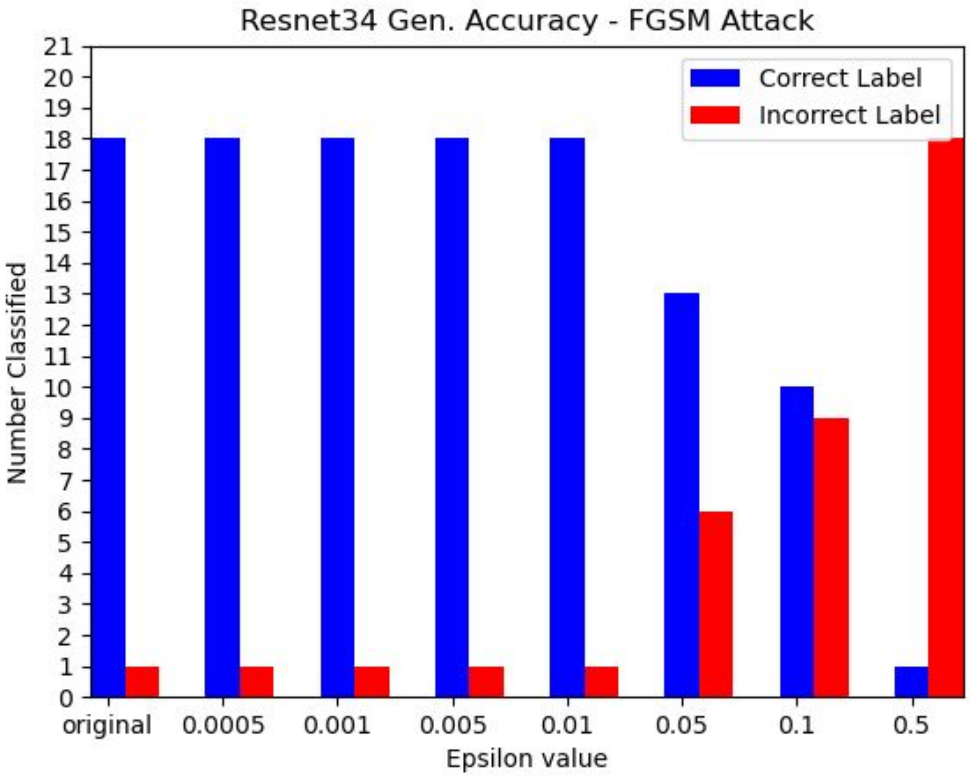
Results - GoogLeNet



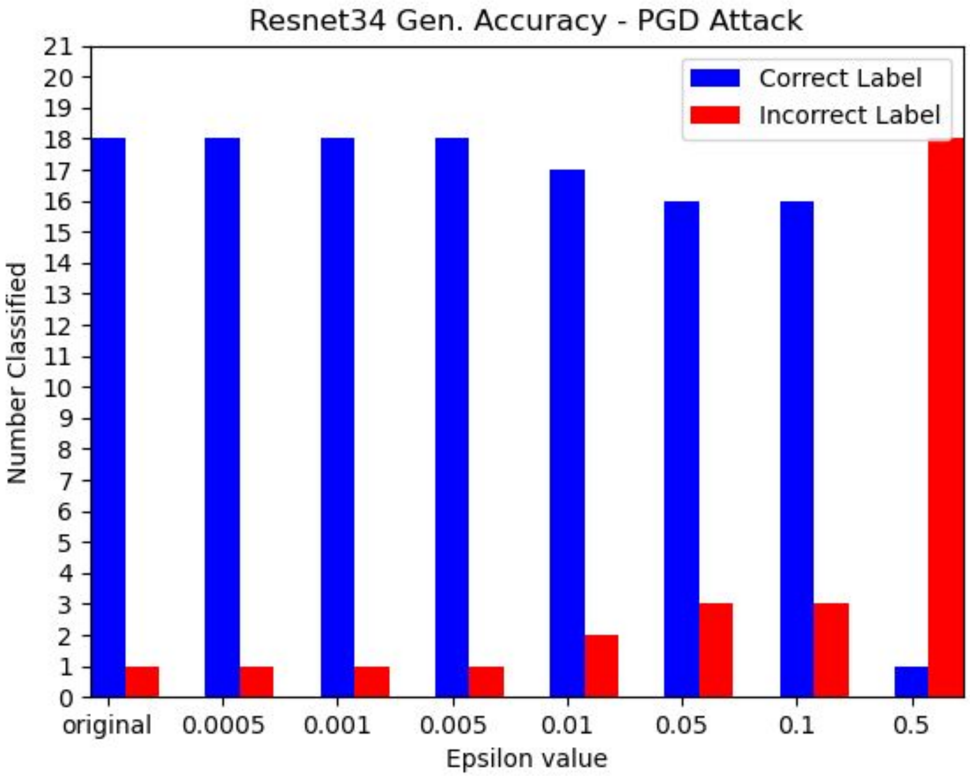
Results - Resnet34



Results - Resnet34



Results - Resnet34



Analysis of Results & Code Review

Best attack - FGSM

Effective epsilon range 0.005 to 0.1

Missed: TRUE: dalmatian, coach dog, carriage dog, dog breed, dog // GIVEN: carnivore or cat or felidae EPSVAL: 0.05

Missed: TRUE: Loafer, shoe // GIVEN: footwear or product or synthetic rubber EPSVAL: 0.05

Analysis of Results & Code Review

Missed: TRUE: bottlecap, bottle cap // GIVEN: font or art or circle EPSVAL: original

Missed: TRUE: cannon // GIVEN: wheel or bicycle tire or bicycle wheel EPSVAL: 0.1

Conclusions - Google Cloud Vision

Effectively robust

Errs on the side of overly general labels if low confidence - imagenet label specificity largely incompatible with Google labels (manual re-labeling required)

Loss of accuracy/misclassification rate correlated with high amount of visible, human-perceptible perturbation

For chosen attacks/models, epsilon 0.005 to 0.01 seems to be sweet spot for trade off of imperceptibility and misclassification

Black-box attacks largely ineffective

Amazon Rekognition

Major Works:

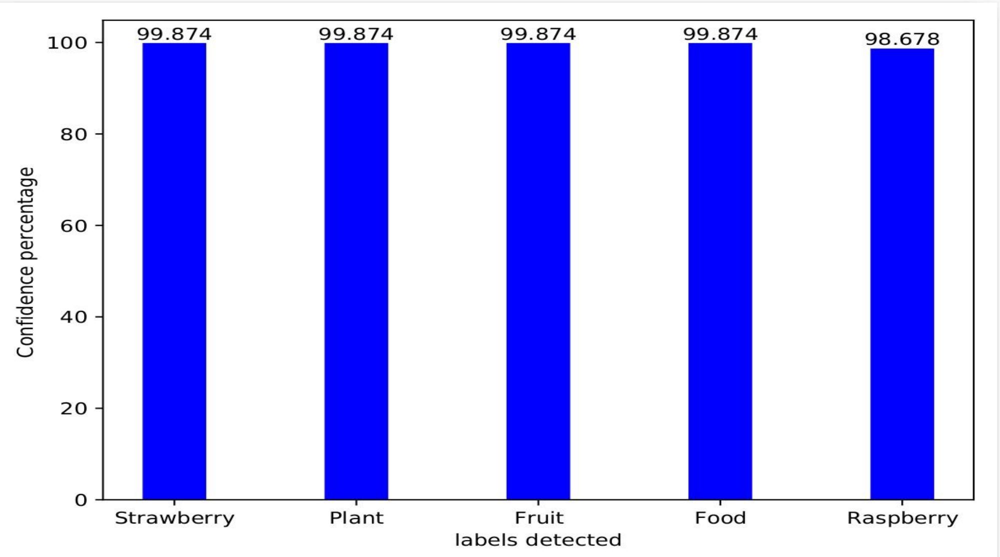
- Implemented API test using both S3 bucket and from the local directory.
- Taken Images (both original and with attacks) with different levels of perturbation ranging from 0.01 to 0.5.
- Checked if the API detected object labels present in the image correctly.
- Displayed the robustness of API using a bar graph showing labels and corresponding confidence percentage.

Experimental Results:

FGSM attack results:

1. Original Image:

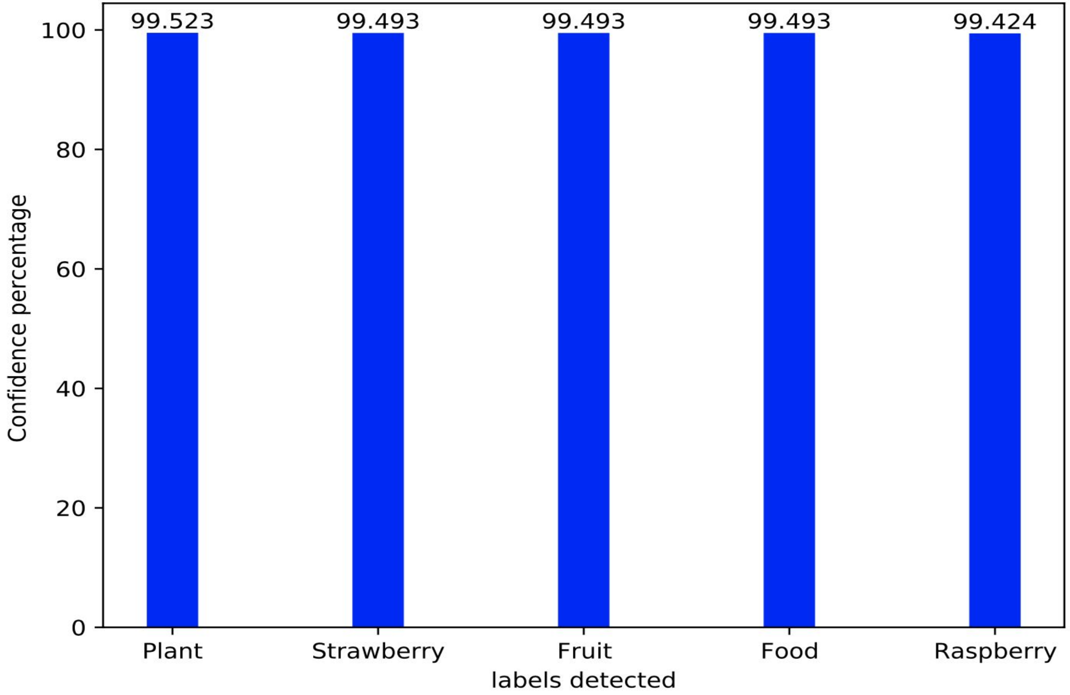
Labels detected by API:



FGSM Continued…

Epsilon = 0.01

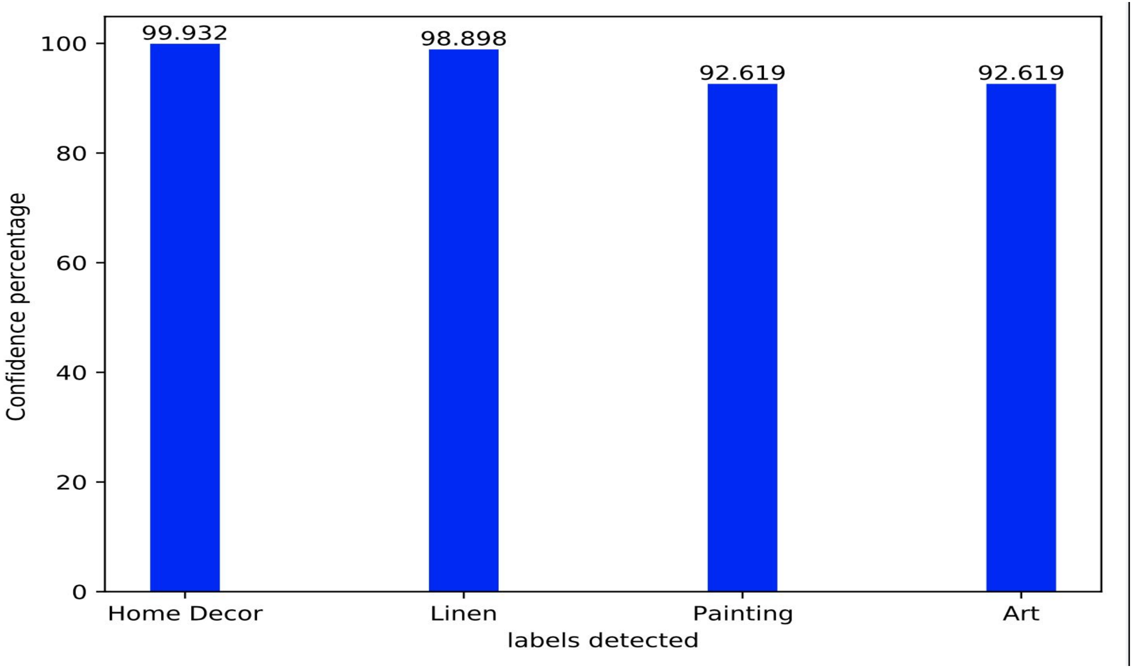
Labels detected by API:



FGSM Continued…

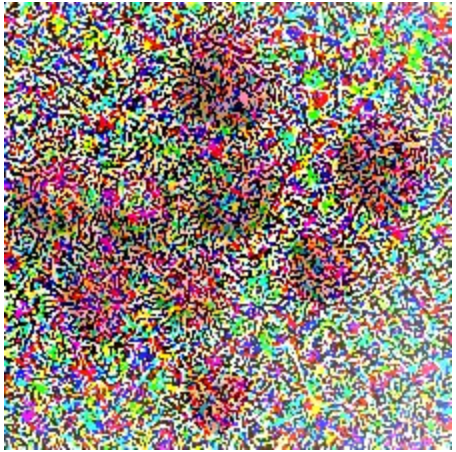
Epsilon = 0.1

Labels detected by API:

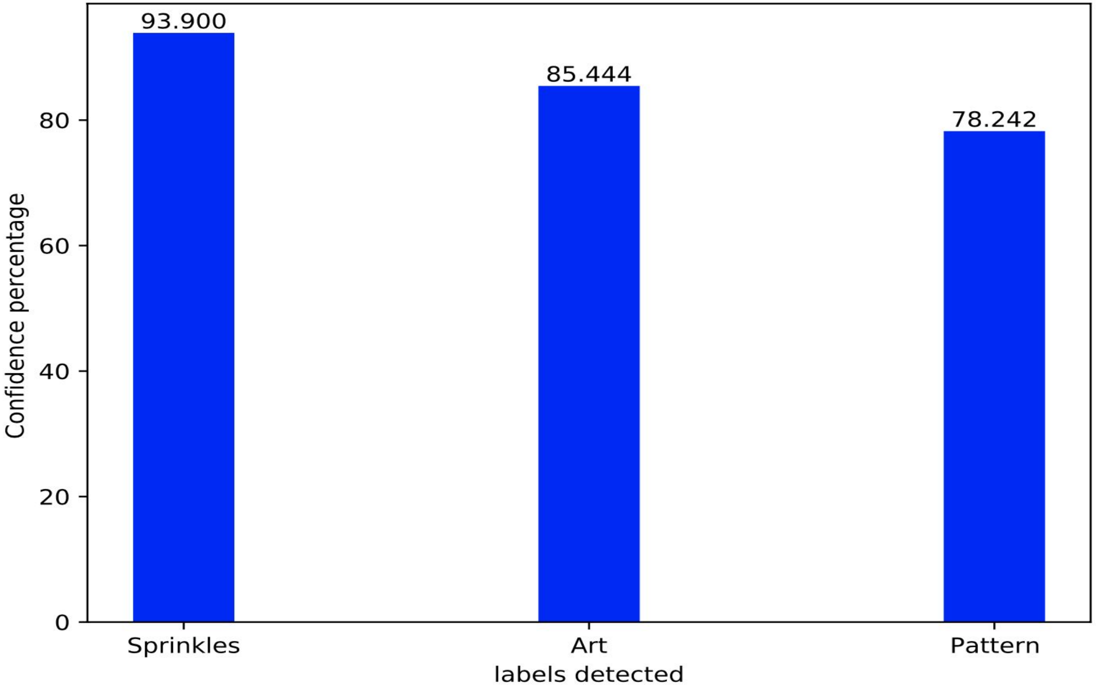


FGSM Continued…

Epsilon = 0.5



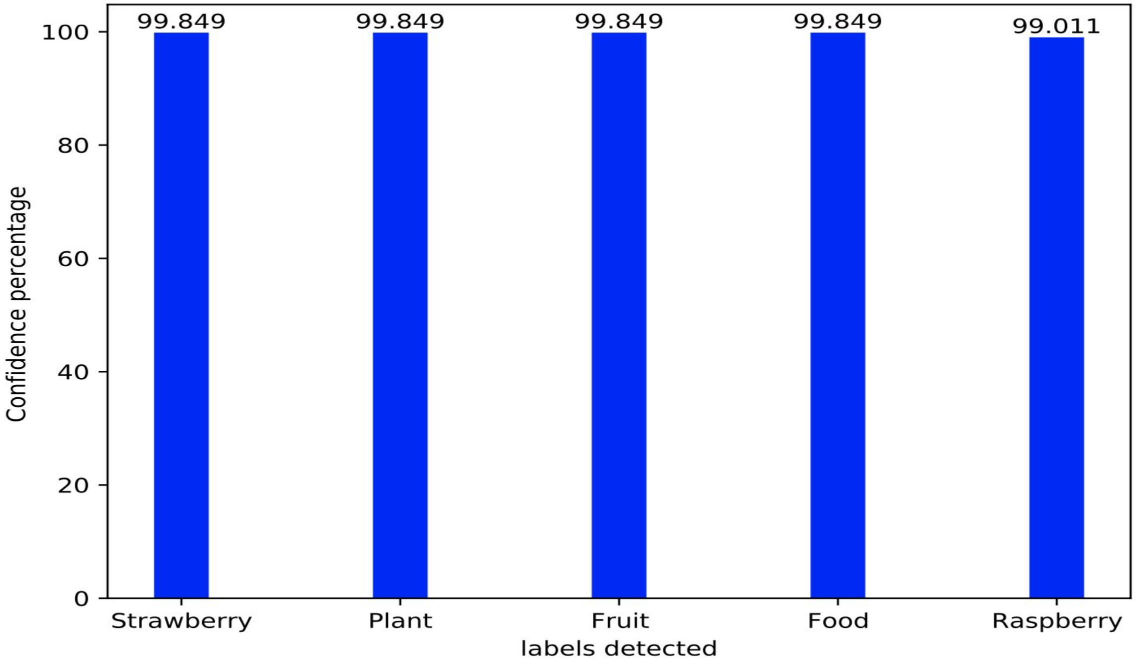
Labels detected by API:



PGD attack Results:

Epsilon = 0.01

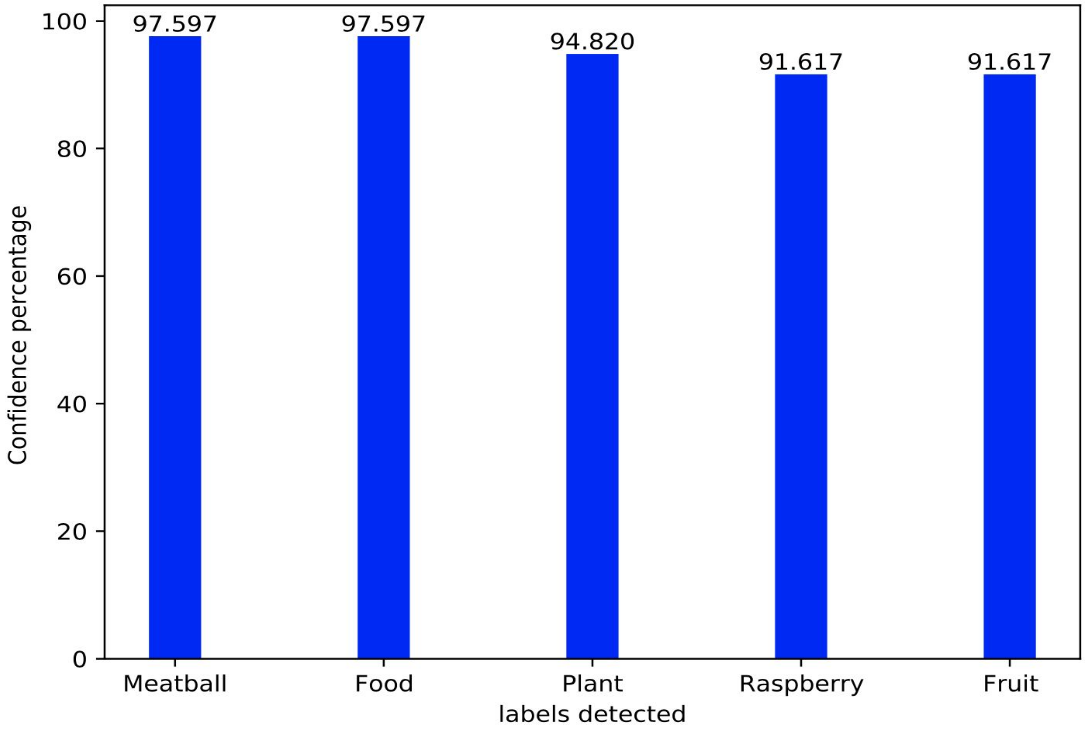
Labels detected by API:



PGD Continued…

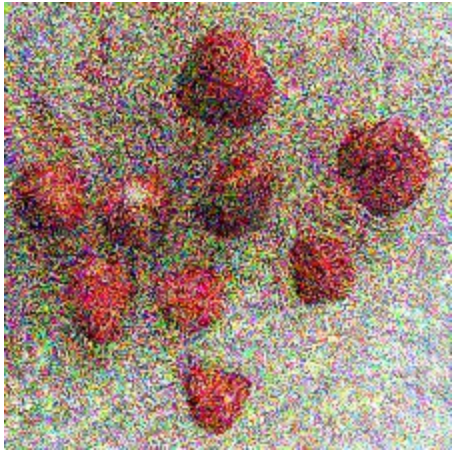
Epsilon = 0.1

Labels detected by API:

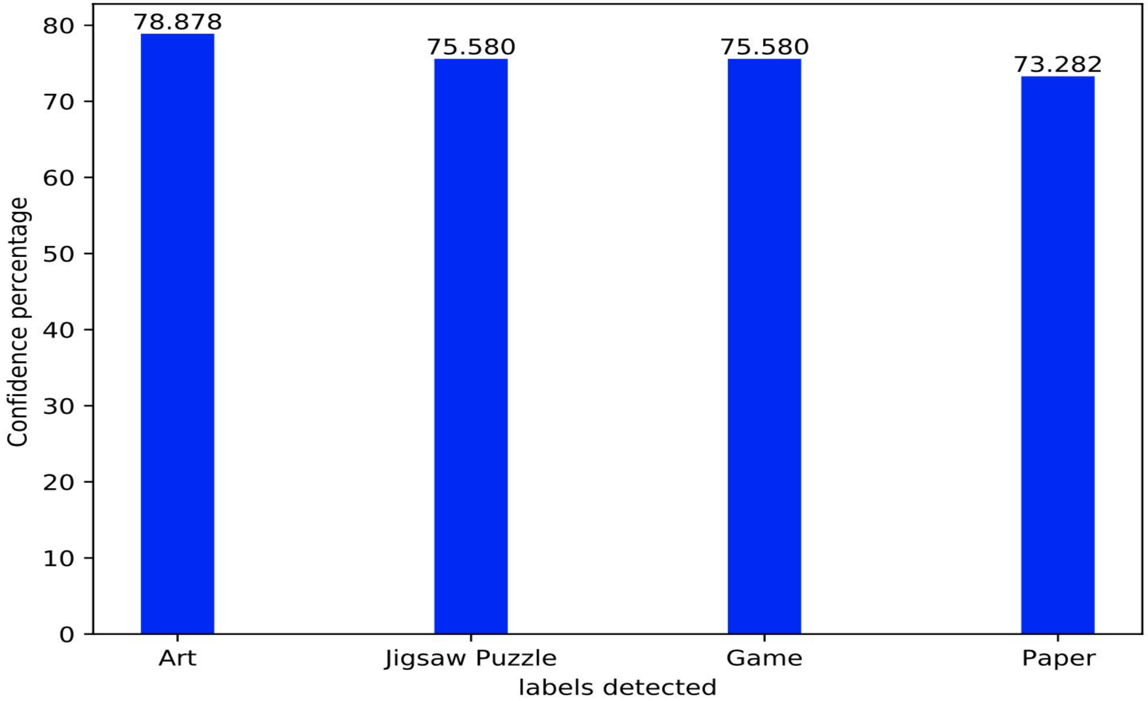


PGD Continued…

Epsilon = 0.5



Labels detected by API:

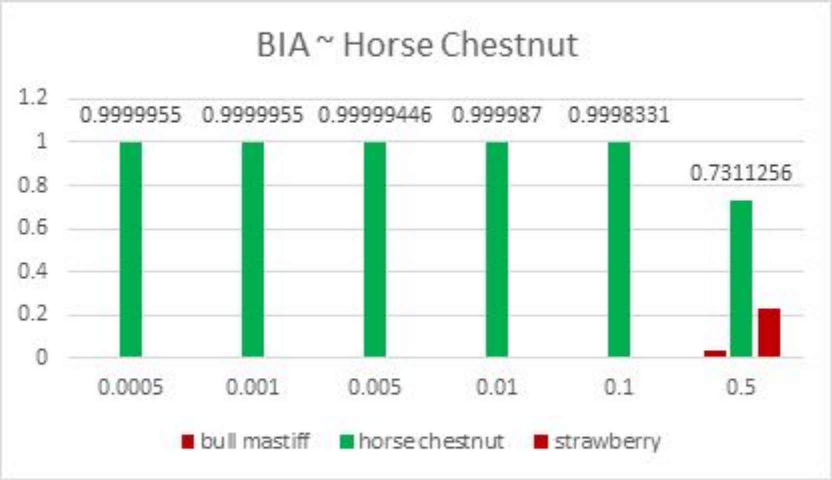


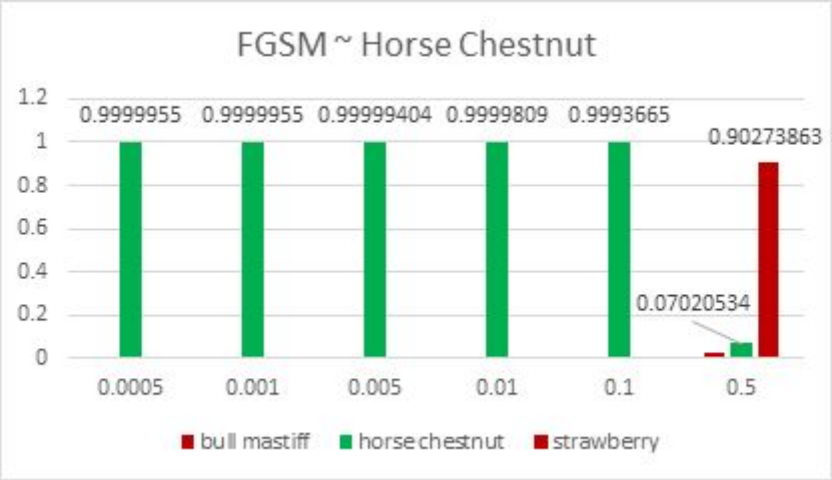
Conclusion

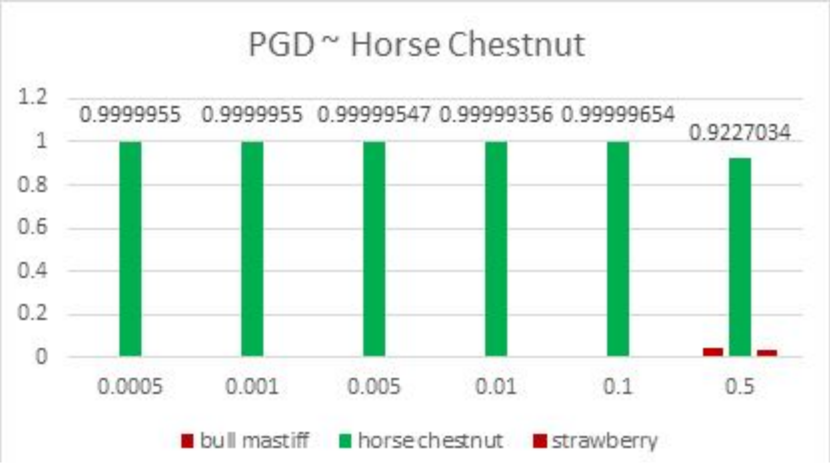
- Successfully tested Amazon Rekognition API using both original image and images with attacks.
- To test the API Adversarial attacks like FGSM and PGD were used.
- For FGSM attack the API failed to detect once the value of epsilon reached 0.1 as shown in the earlier confidence bar graph.
- For PGD model the API again failed once epsilon was 0.1 but API performed slightly better than with FGSM.

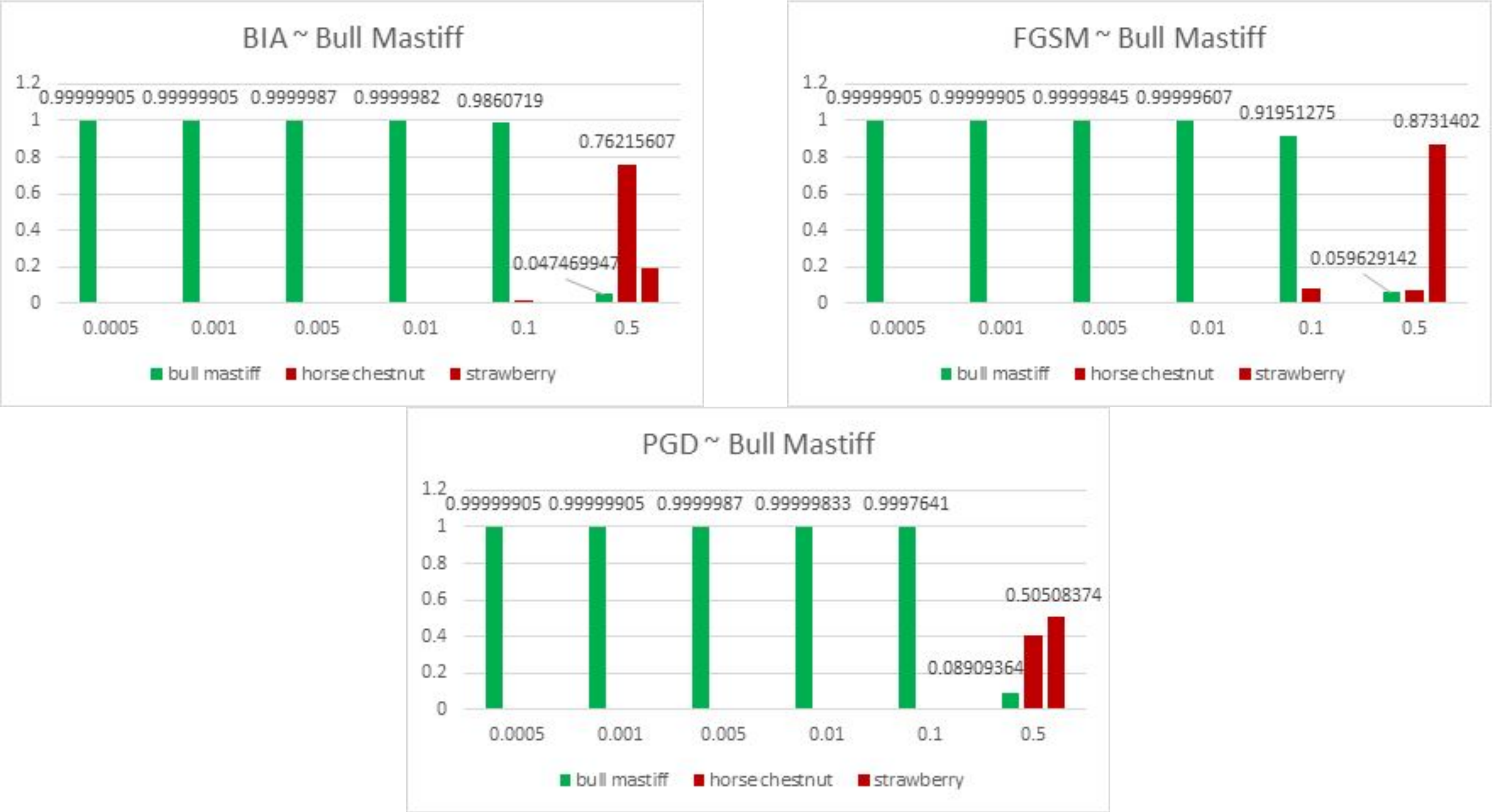
Salesforce MetaMind Einstein Vision

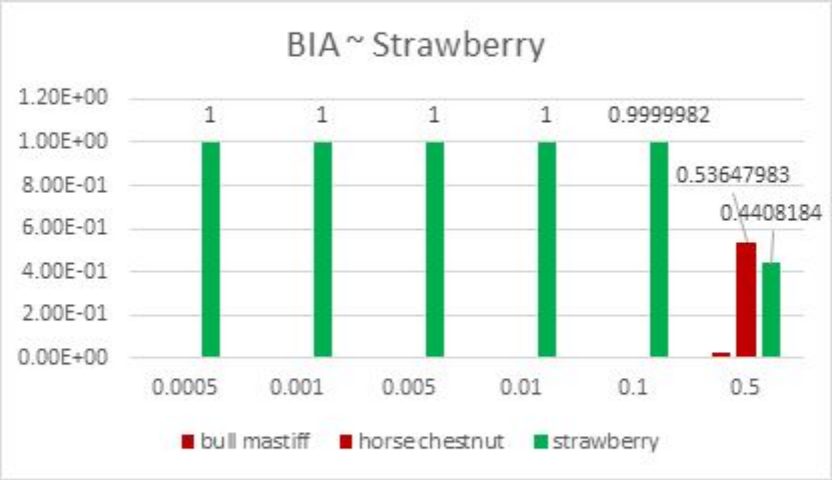
- Documentation: https://metamind.readme.io/docs
- Using the Einstein API, an image custom classifier was created
- API uses cUrl and an access token that is granted to a user with a .pem file
- Image classifier and model fairly restrictive in comparison to previous API’s
- Used images pre-trained by Resnet34 perturbed using BIA, FGSM, and PGD

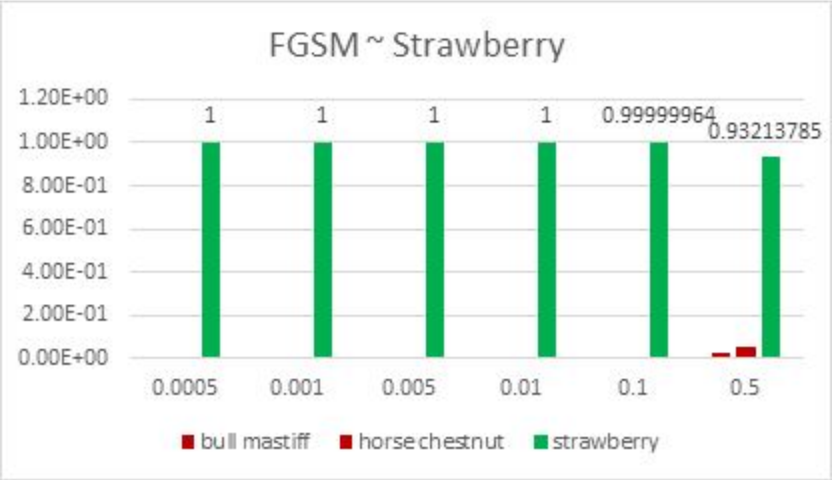

BIA ~ Horse Chestnut
1.2
1
0.8
0.6
0.4
0.2
0
0.9999955
0.9999955
0.99999446
0.999987
0.9998331
0.7311256
0.0005
0.001
0.005
0.01
0.1
0.5
bull mastiff
horse chestnut
strawberry

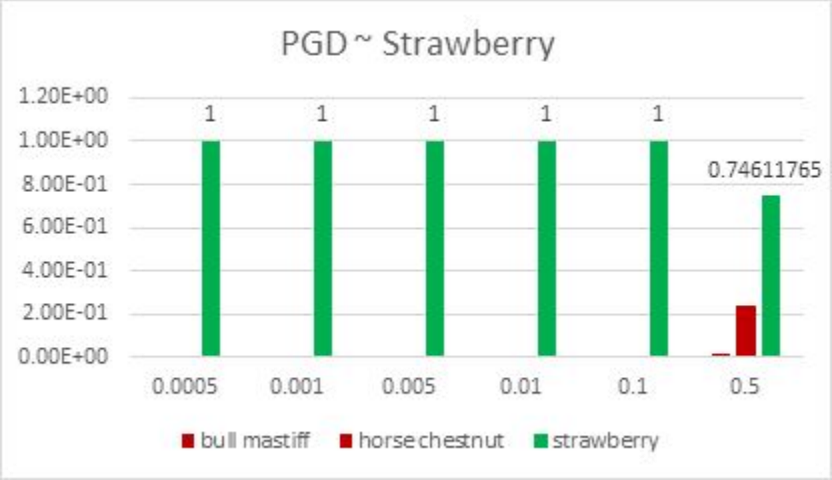

FGSM ~ Horse Chestnut
1.2
1
0.8
0.6
0.4
0.2
0
0.9999955
0.9999955
0.99999404
0.9999809
0.9993665
0.90273863
0.07020534
0.0005
0.001
0.005
0.01
0.1
0.5
bull mastiff
horse chestnut
strawberry

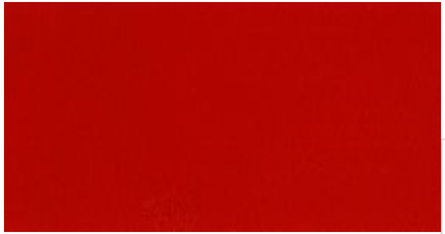

PGD ~ Horse Chestnut
1.2
1
0.8
0.6
0.4
0.2
0
0.9999955
0.9999955
0.99999547
0.99999356
0.99999654
0.9227034
0.0005
0.001
0.005
0.01
0.1
0.5
bull mastiff
horse chestnut
strawberry

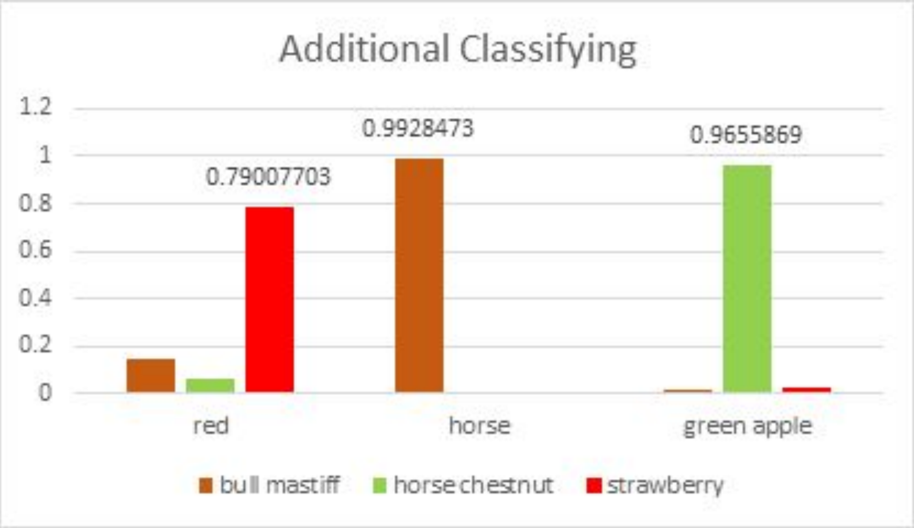

BIA ~ Bull Mastiff
1.2
1
0.8
0.6
0.4
0.2
0
0.99999905 0.99999905 0.9999987 0.9999982 0.9860719
0.76215607
0.047469947
0.0005 0.001 0.005 0.01 0.1 0.5
bull mastiff horse chestnut strawberry
FGSM ~ Bull Mastiff
1.2
1
0.8
0.6
0.4
0.2
0
0.99999905 0.99999905 0.99999845 0.99999607
0.91951275
0.8731402
0.059629142
0.0005 0.001 0.005 0.01 0.1 0.5
bull mastiff horse chestnut strawberry
PGD ~ Bull Mastiff
1.2
1
0.8
0.6
0.4
0.2
0
0.99999905 0.99999905 0.9999987 0.99999833 0.9997641
0.50508374
0.08909364
0.0005 0.001 0.005 0.01 0.1 0.5
bull mastiff horse chestnut strawberry


BIA ~ Strawberry
1.20E+00
1.00E+00
8.00E-01
6.00E-01
4.00E-01
2.00E-01
0.00E+00
1
1
1
1
0.9999982
0.53647983
0.4408184
0.0005
0.001
0.005
0.01
0.1
0.5
bull mastiff
horse chestnut
strawberry


FGSM ~ Strawberry
1.20E+00
1.00E+00
8.00E-01
6.00E-01
4.00E-01
2.00E-01
0.00E+00
1
1
1
1
0.99999964
0.93213785
0.0005
0.001
0.005
0.01
0.1
0.5
bull mastiff
horse chestnut
strawberry


PGD ~ Strawberry
1.20E+00
1.00E+00
8.00E-01
6.00E-01
4.00E-01
2.00E-01
0.00E+00
1
1
1
1
1
0.74611765
0.0005
0.001
0.005
0.01
0.1
0.5
bull mastiff
horse chestnut
strawberry




Additional Classifying
1.2
1
0.8
0.6
0.4
0.2
0
0.79007703
0.9928473
0.9655869
red
horse
green apple
bull mastiff
horse chestnut
strawberry

Conclusion

- Robust but restrictive
- Black-box untargeted attacks ineffective; targeted attacks possibly effective
- Unconventional ways to attack model
- Future Work - modify dataset and continue to experiment

Questions