SimBA: Black box attacks on Image classifiers

Team 33: Ardhendu Banerjee (2022201005)
Animesh Das (2022201027)
Ritvik Gupta (2022202005)

Paper Link: Simple Black-box Adversarial Attacks (arxiv.org)

The problem

Paper

Simple Black Box adversarial attacks
Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, Kilian Q. Weinberger. 2019.

Problem Statement

Try to make pretrained state of the art classifiers on Imagenet to misclassify data, using as less queries as possible in a black box fashion.

Experiment Setup

Experiments on a subset of Imagenet data (Imagenette) and Tiny Imagenet data with models available through the PyTorch API.

Proposed Solution of the paper

Random Attack

In each attack random pixels are chosen and a channel(color) is chosen at random.

Orthonormal Basis

Now value of a chosen channel is changed in such a way so as to propagate the attack (cartesian basis).

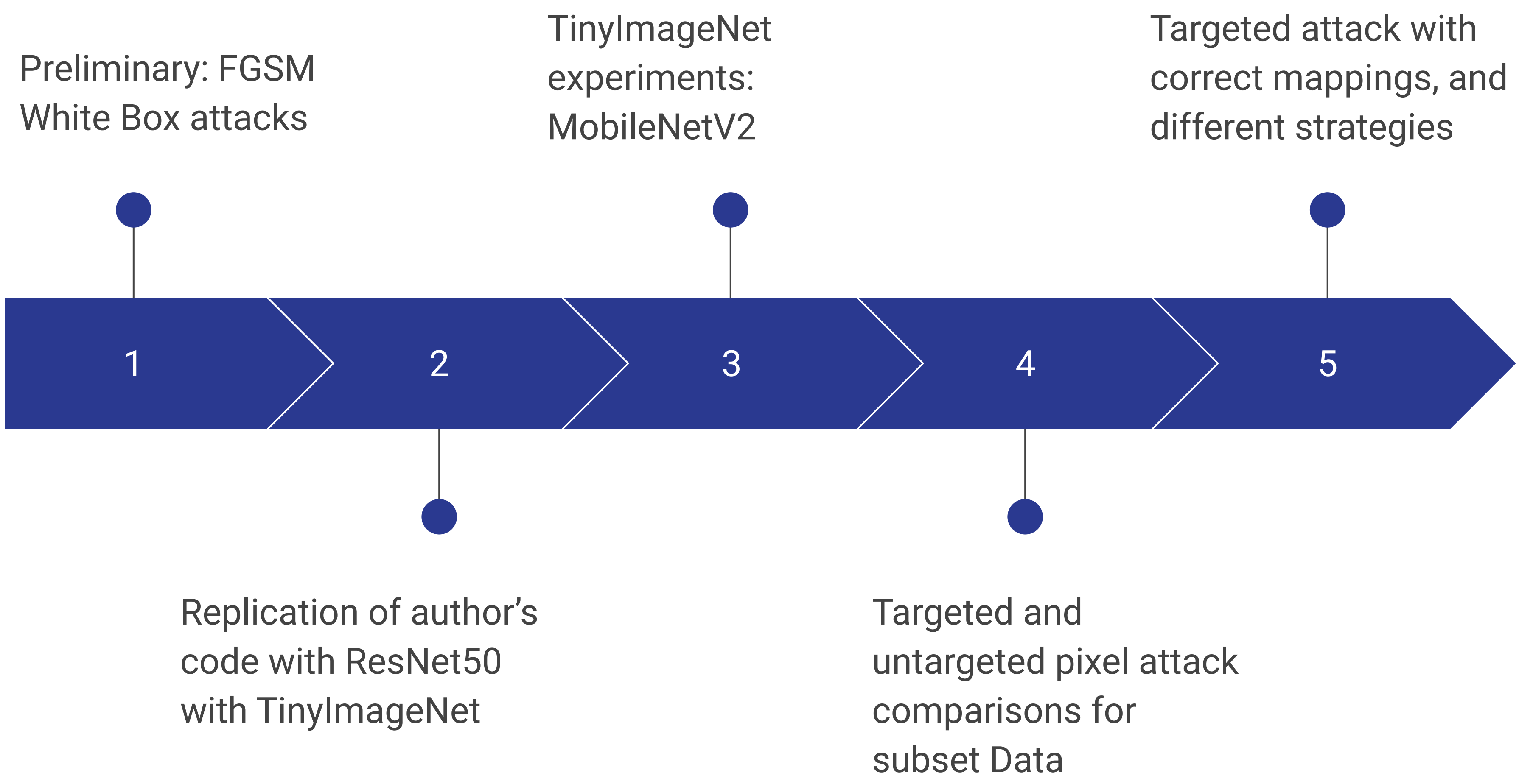
Query Efficiency

Most of the images are misclassified successfully after only 5000 queries (70% success rate, as reported)

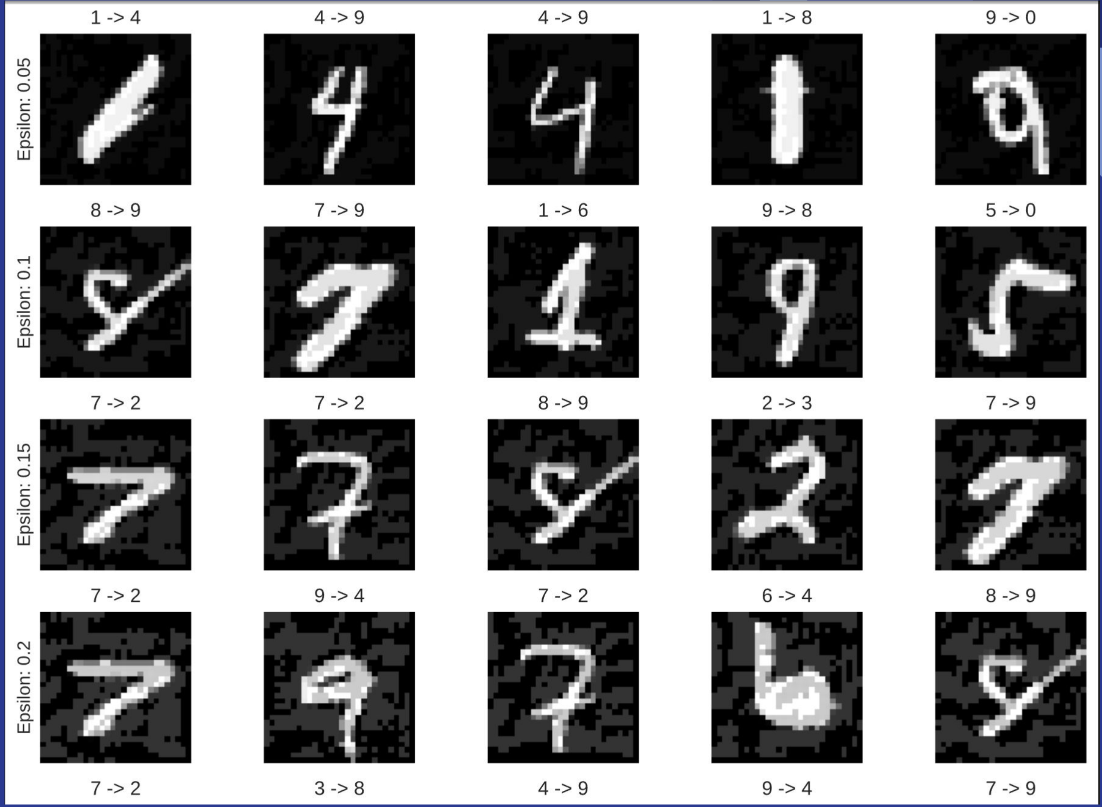
Attack Strategy

Cartesian Basis: a single independent feature of the sample

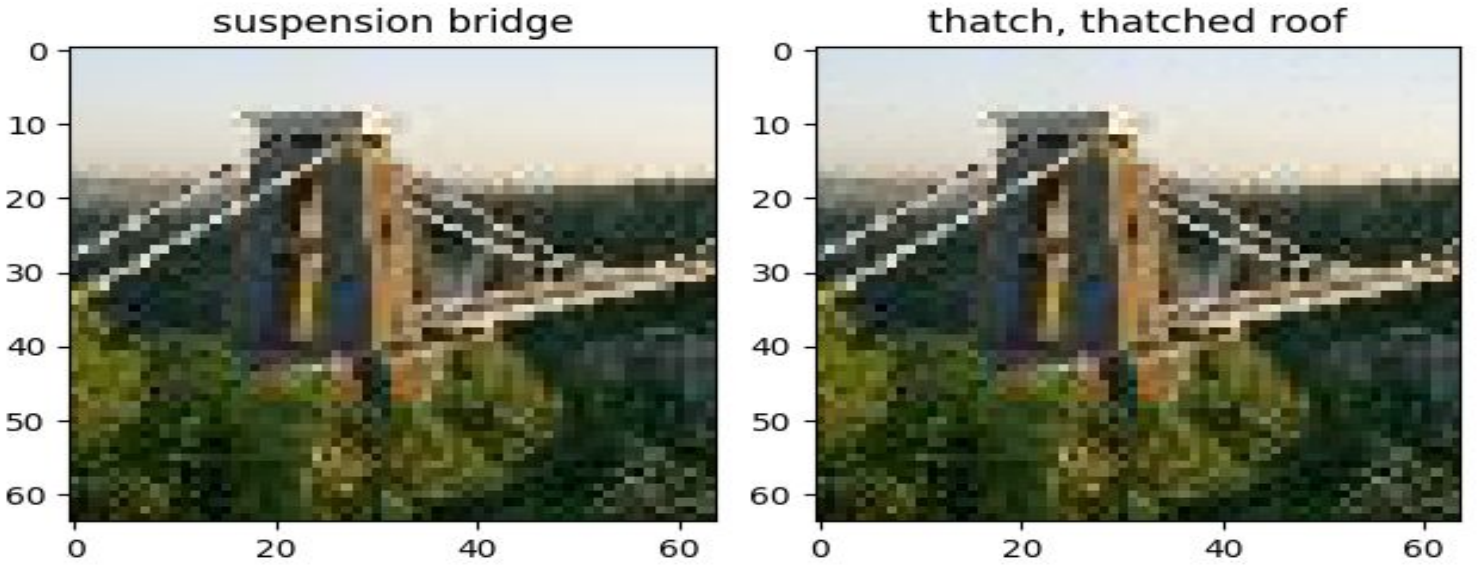
We perturb the individual channels of a pixel in each iteration by a small amount epsilon. At first we add the value to a random pixel, if the model probability decreases, we set that value else subtract from the value and try again. If the probability still doesn’t change, we choose another random pixel in the next iteration.


Preliminary: FGSM White Box attacks
TinyImageNet experiments: MobileNetV2
Targeted attack with correct mappings, and different strategies
1
2
3
4
5
Replication of author’s code with ResNet50 with TinyImageNet
Targeted and untargeted pixel attack comparisons for subset Data

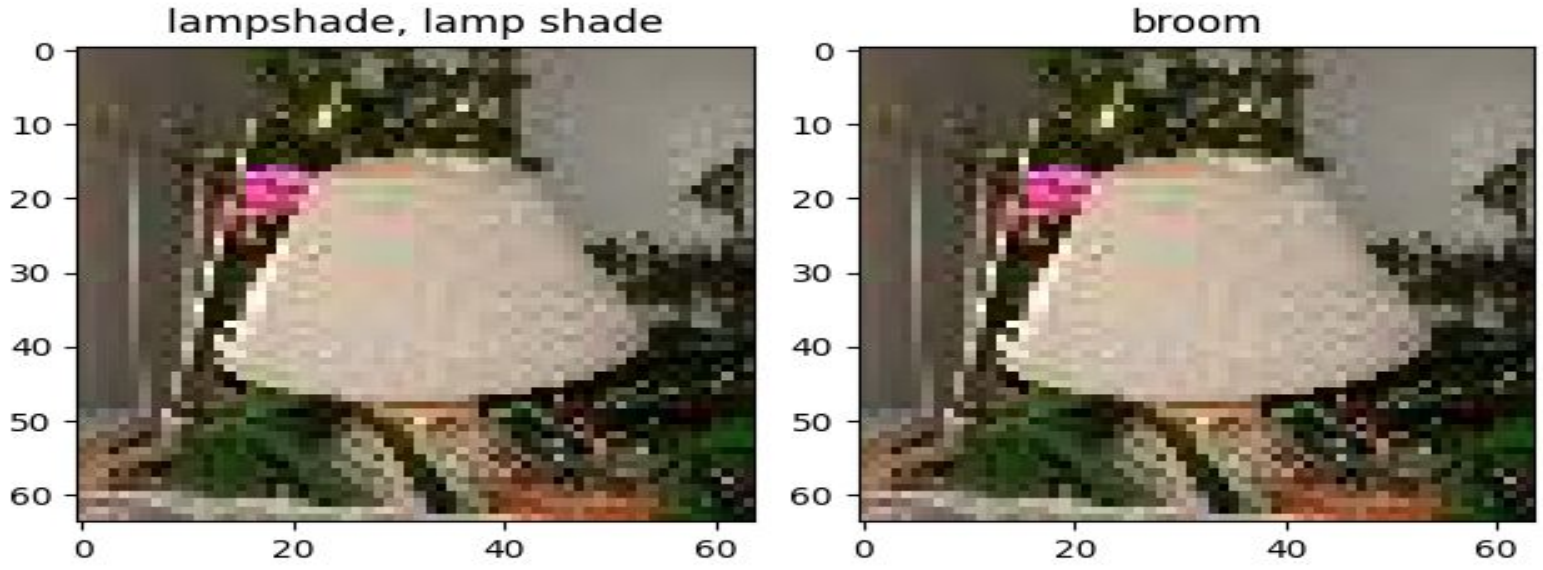
Results


1 -> 4
4 -> 9
4 -> 9
1 -> 8
9 -> 0
Epsilon: 0.05
8 -> 9
7 -> 9
1 -> 6
9 -> 8
5 -> 0
Epsilon: 0.1
7 -> 2
7 -> 2
8 -> 9
2 -> 3
7 -> 9
Epsilon: 0.15
7 -> 2
9 -> 4
7 -> 2
6 -> 4
8 -> 9
Epsilon: 0.2
7 -> 2
3 -> 8
4 -> 9
9 -> 4
7 -> 9

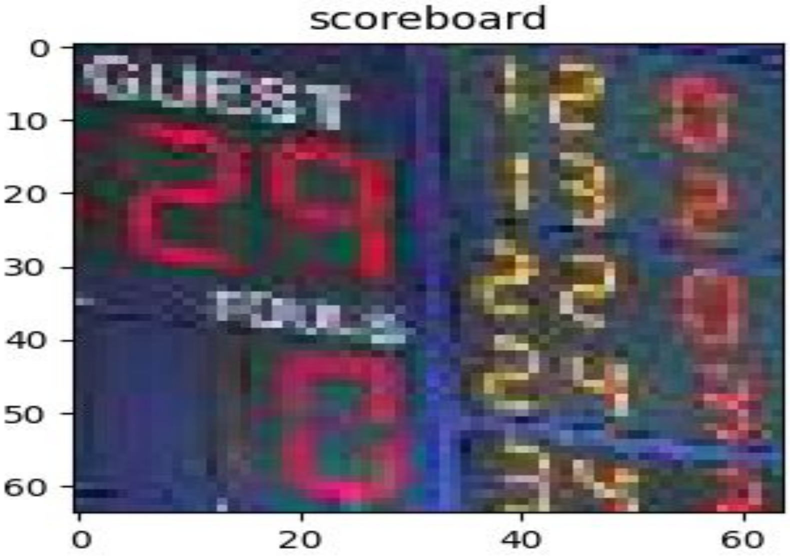
Untargeted attack on Tinyimagenet



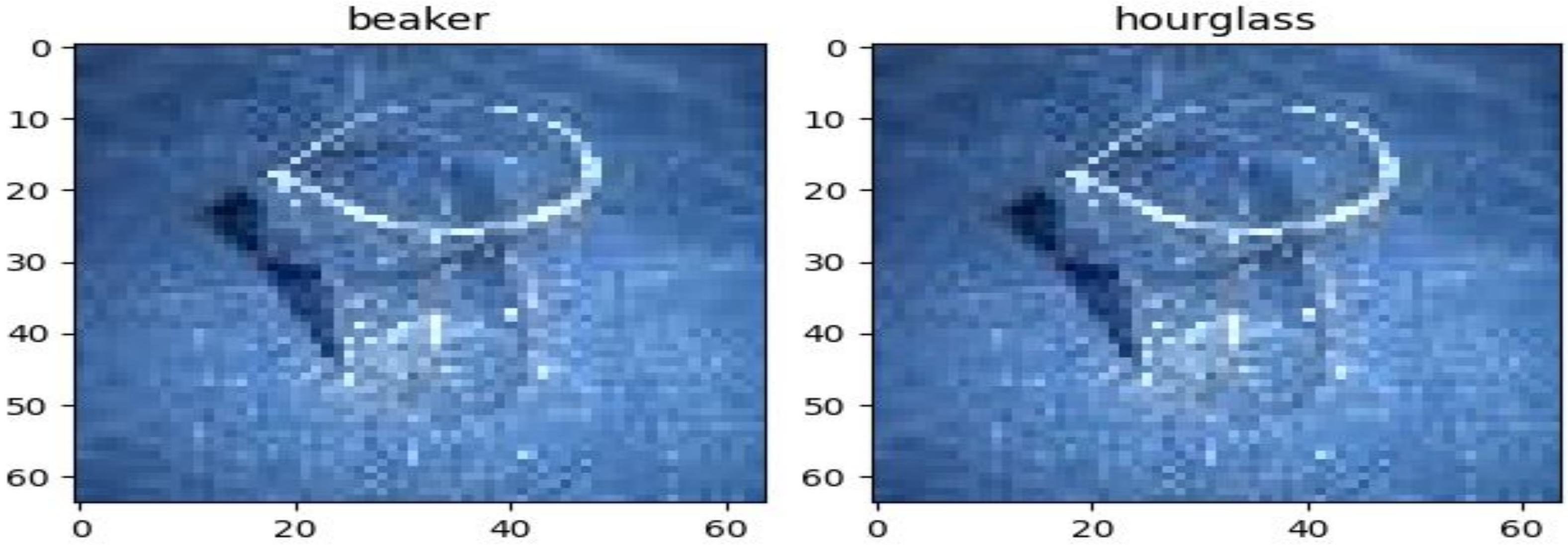
Untargeted attack on Tinyimagenet



Targeted attack on Tinyimagenet: 5th likely class

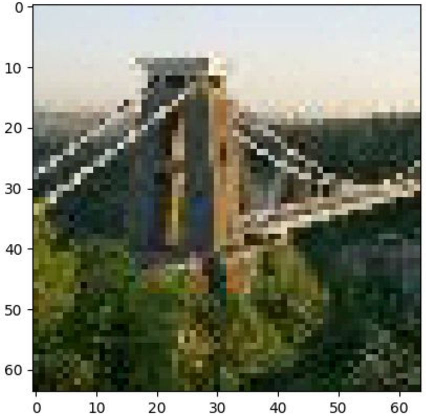


Targeted attack on Tinyimagenet

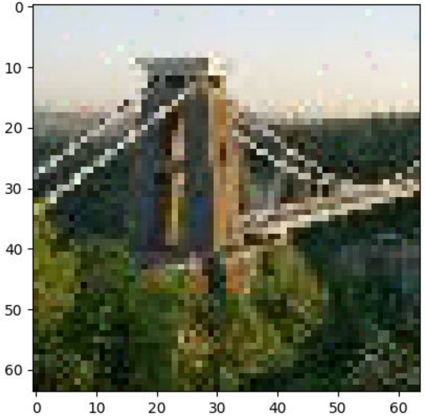


Visibility of attack with increasing Epsilon

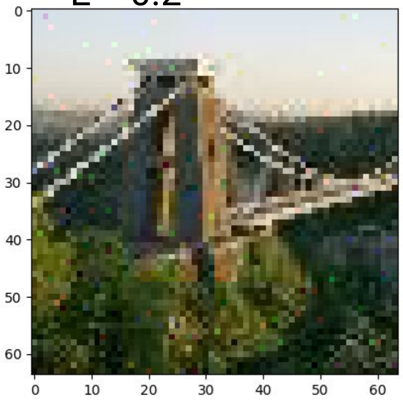
E = 0.02



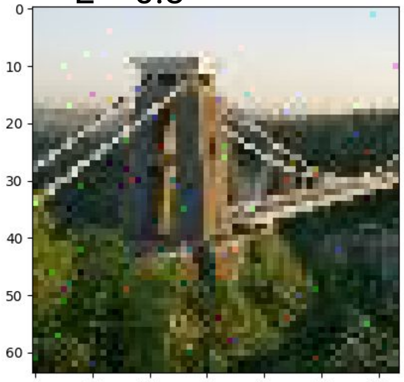
E = 0.1



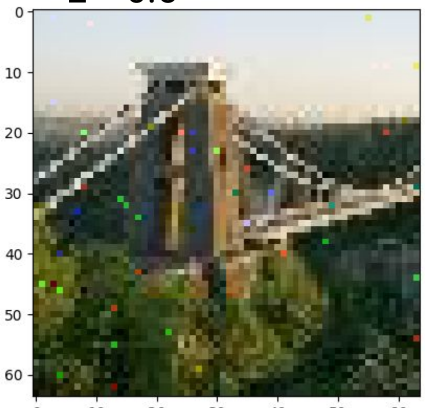
E = 0.2



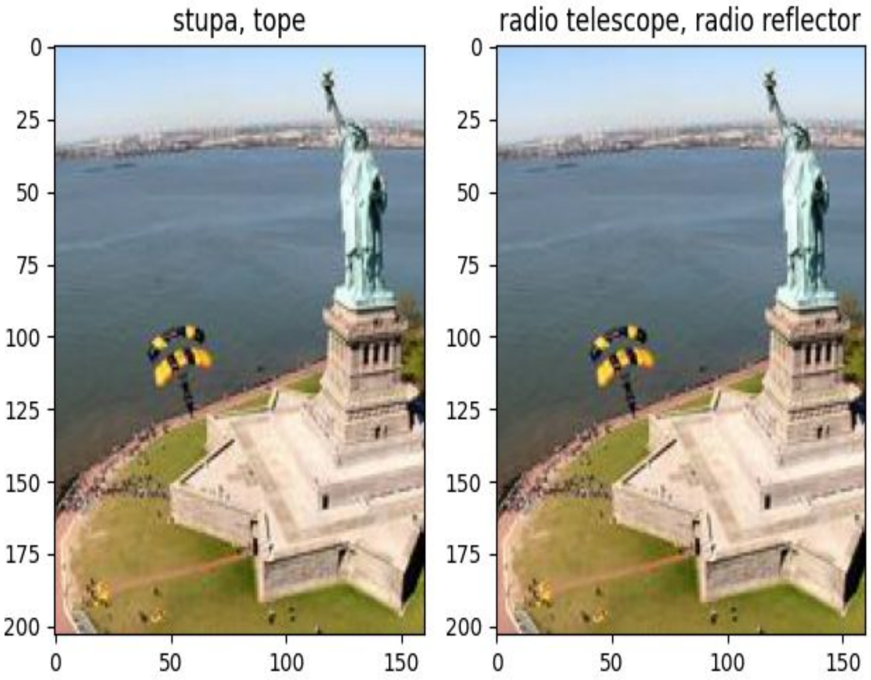
E = 0.3

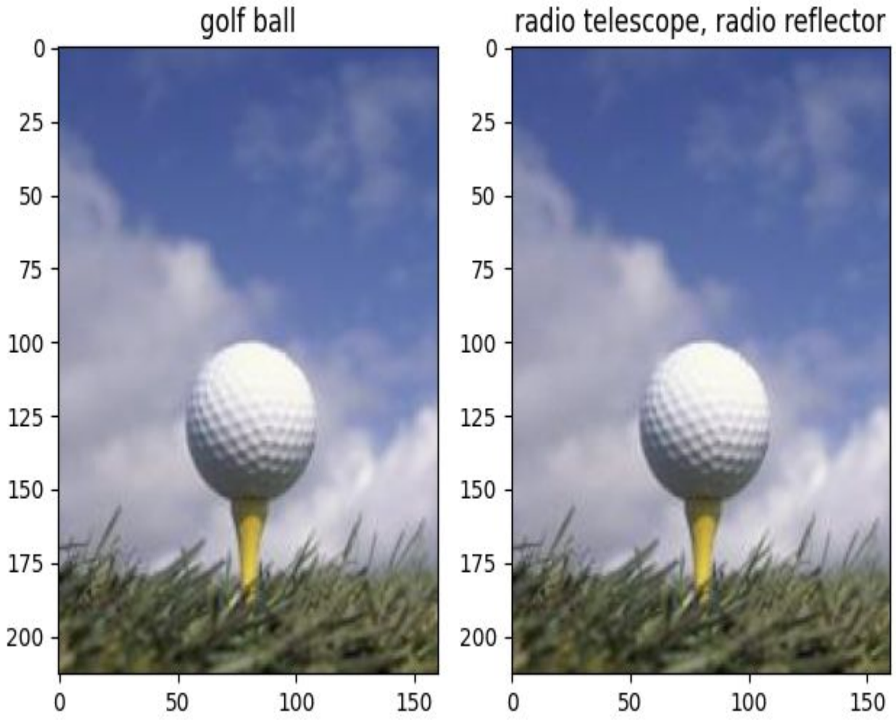


E = 0.5

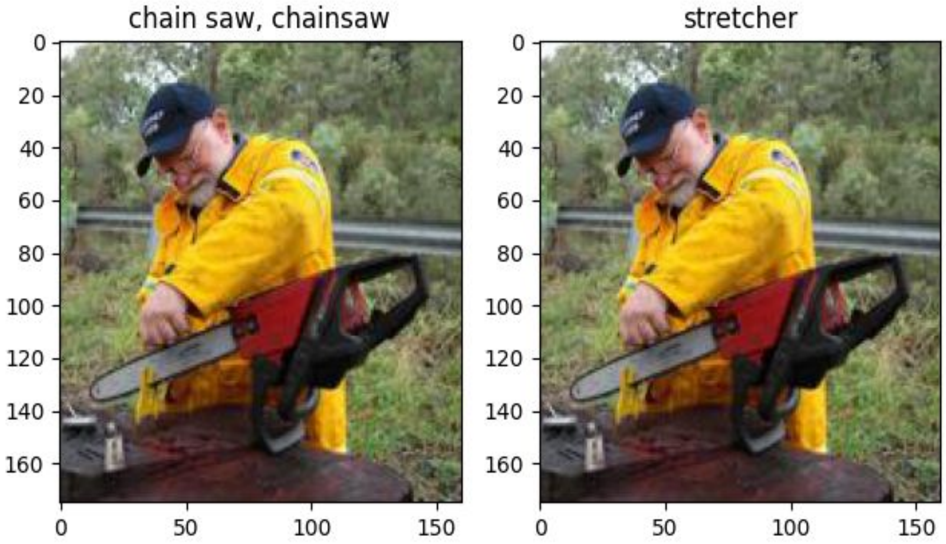


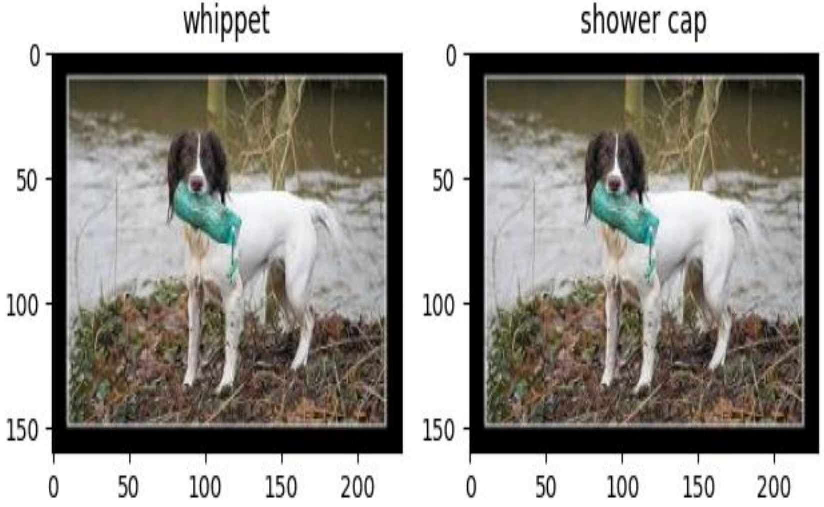
Untargeted attack on subset



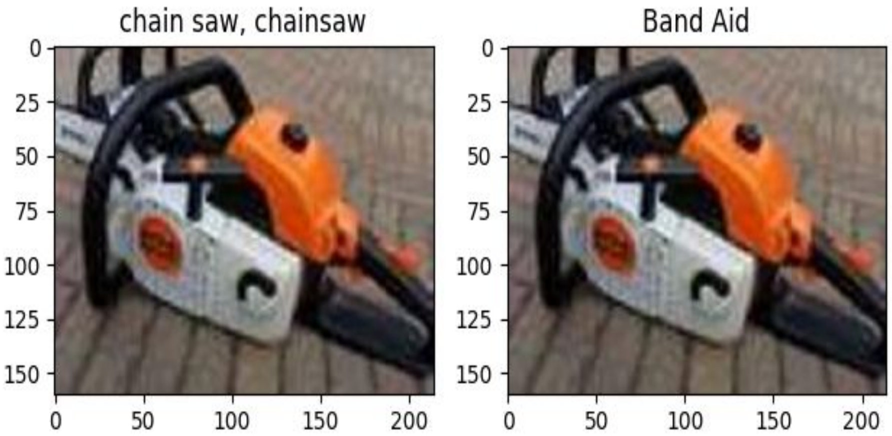


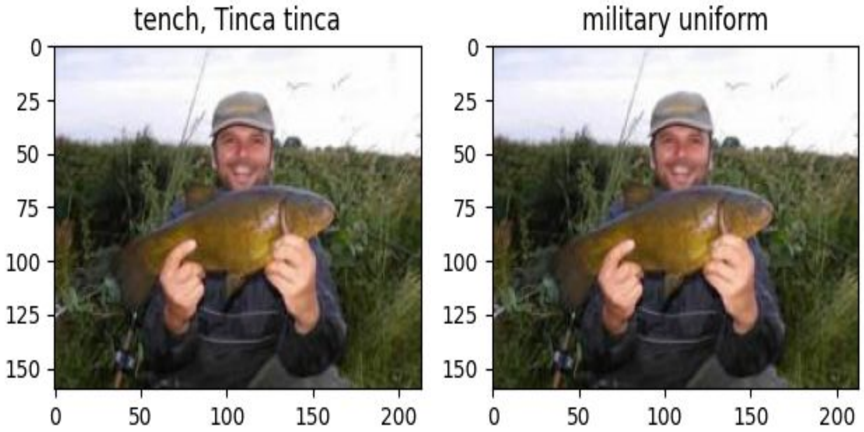
Untargeted attack on subset



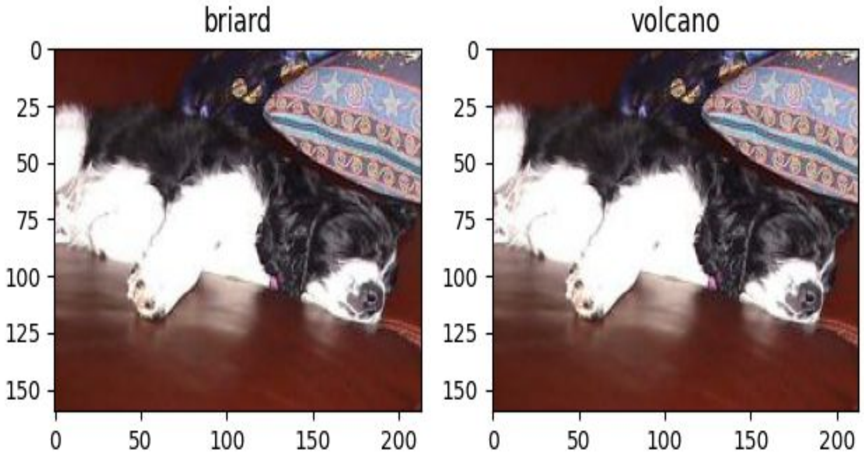


Target attack (epsilon=0.02) on subset

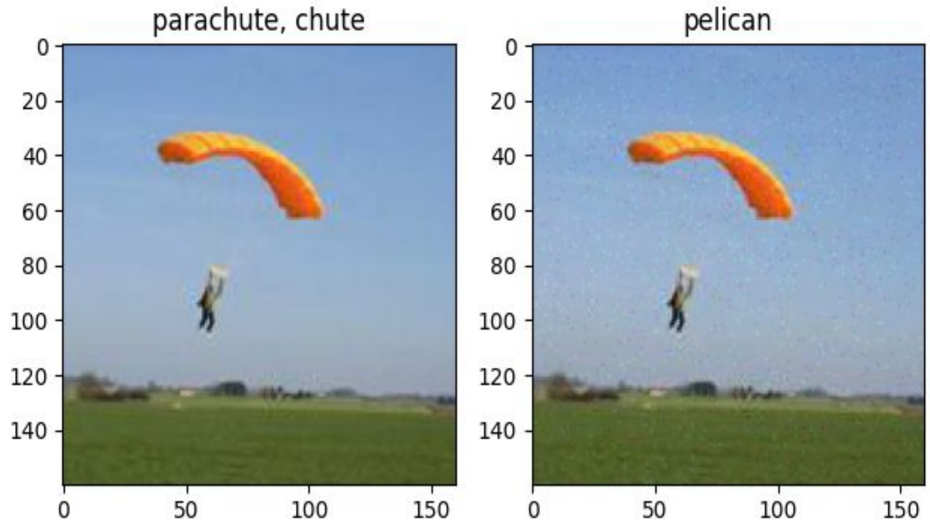




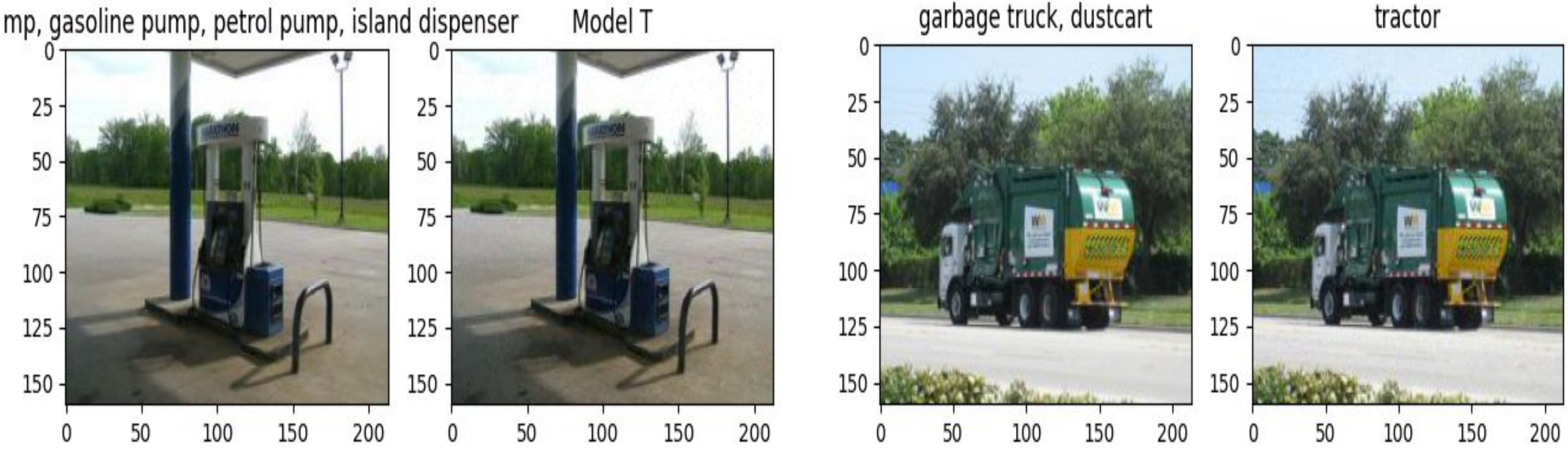
Target attack (epsilon=0.02) on subset



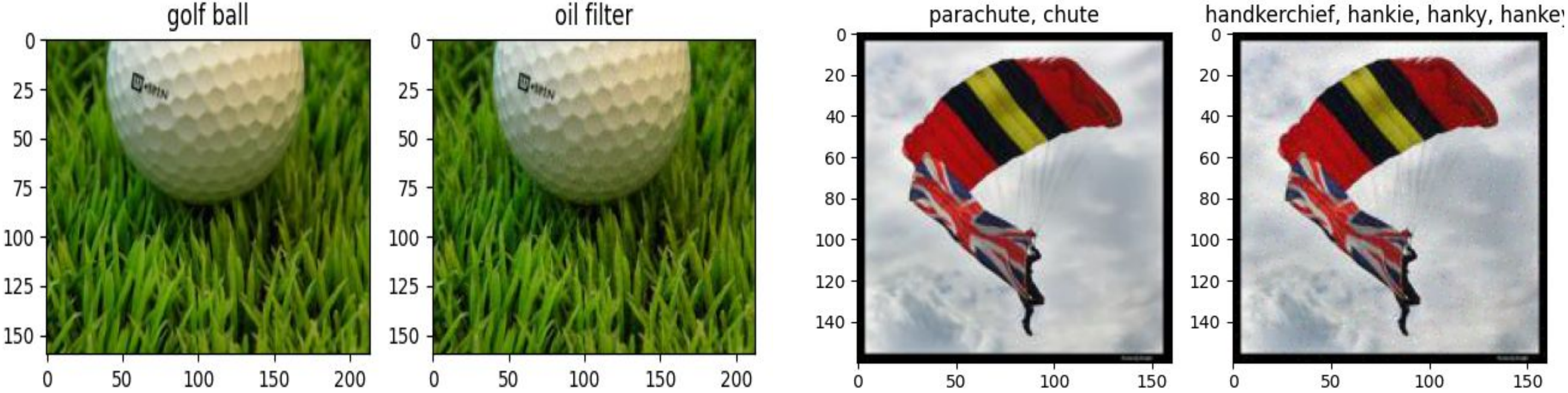
Target attack (epsilon=0.1) on subset



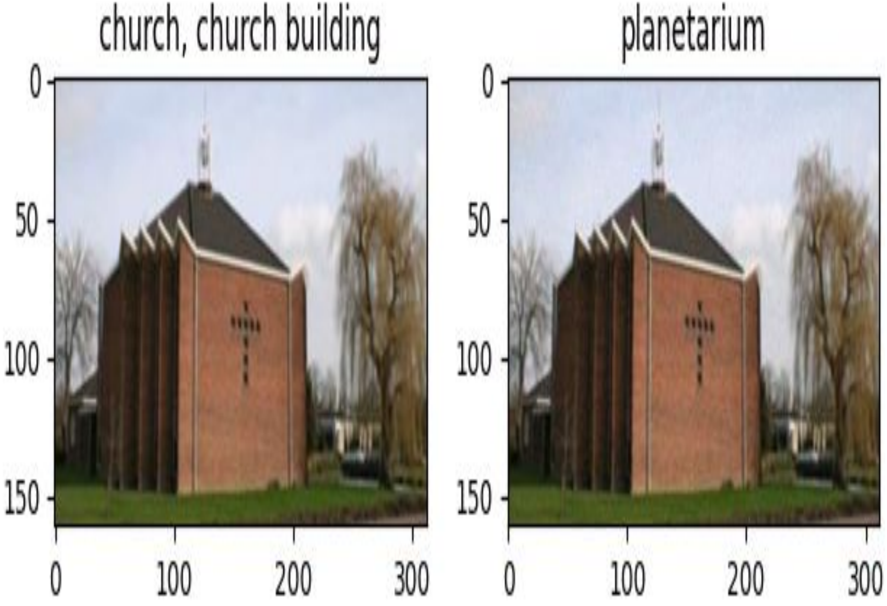
Target attack (epsilon=0.1) on subset

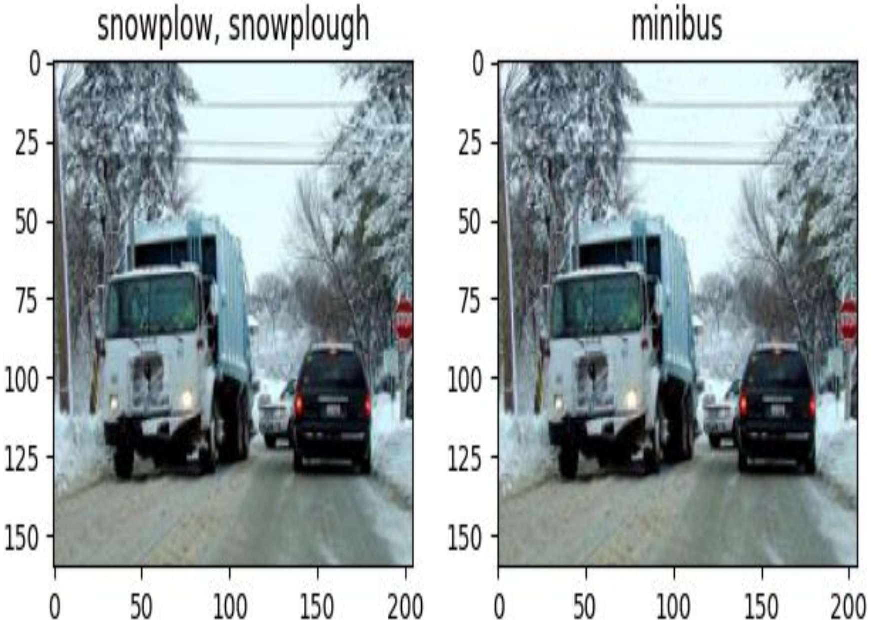


Target attack (epsilon=0.1) on subset

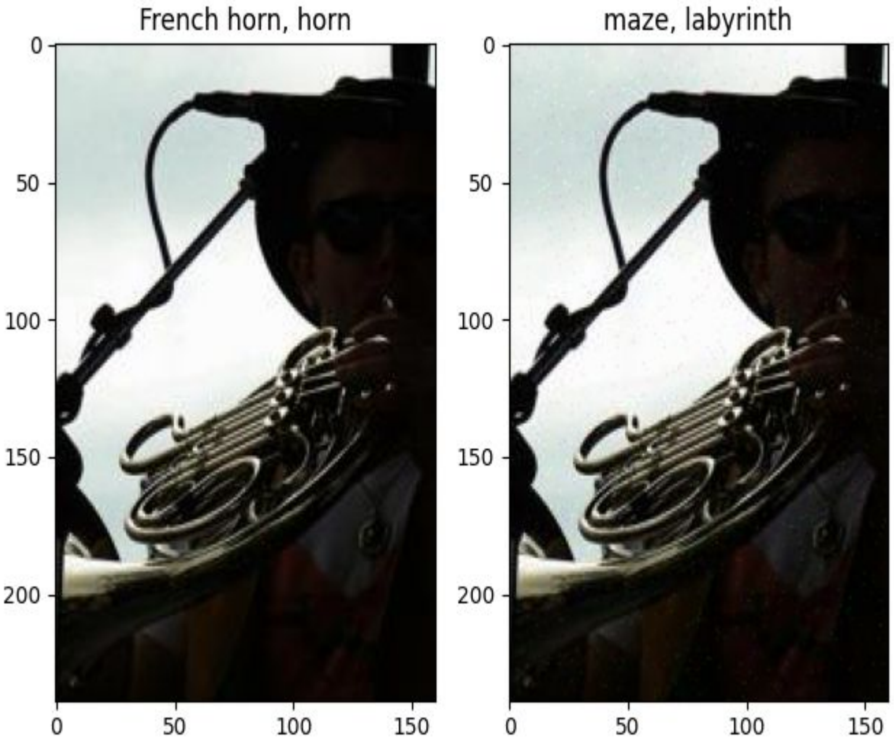


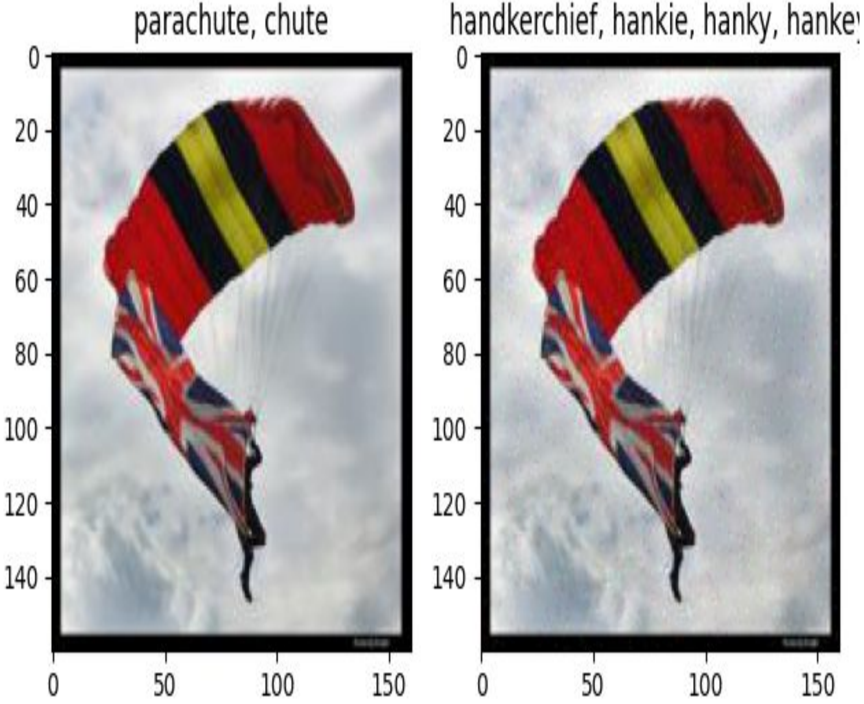
Target attack (epsilon=0.2) on subset



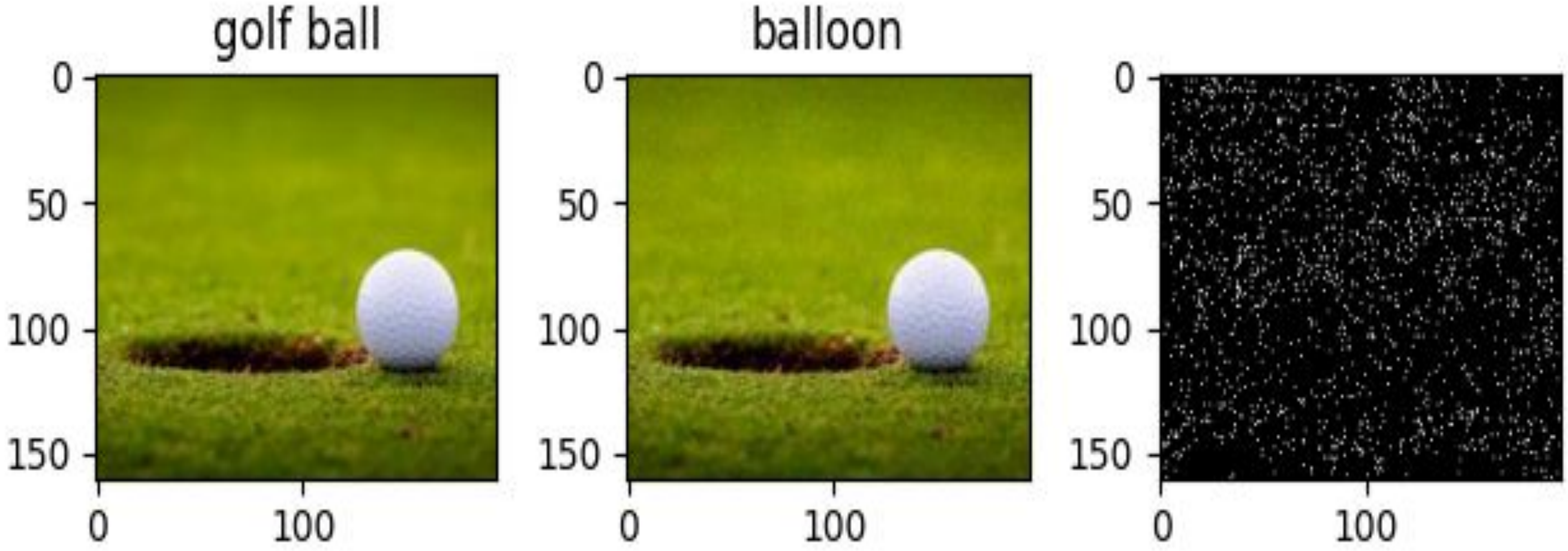


Target attack (epsilon=0.2) on subset

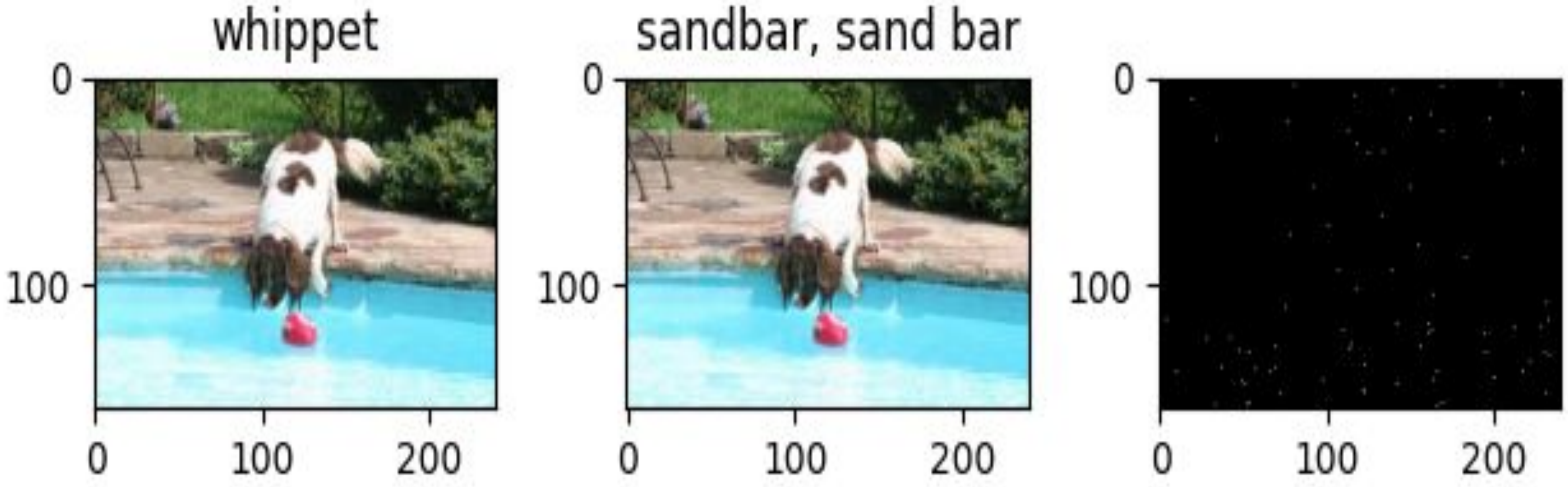




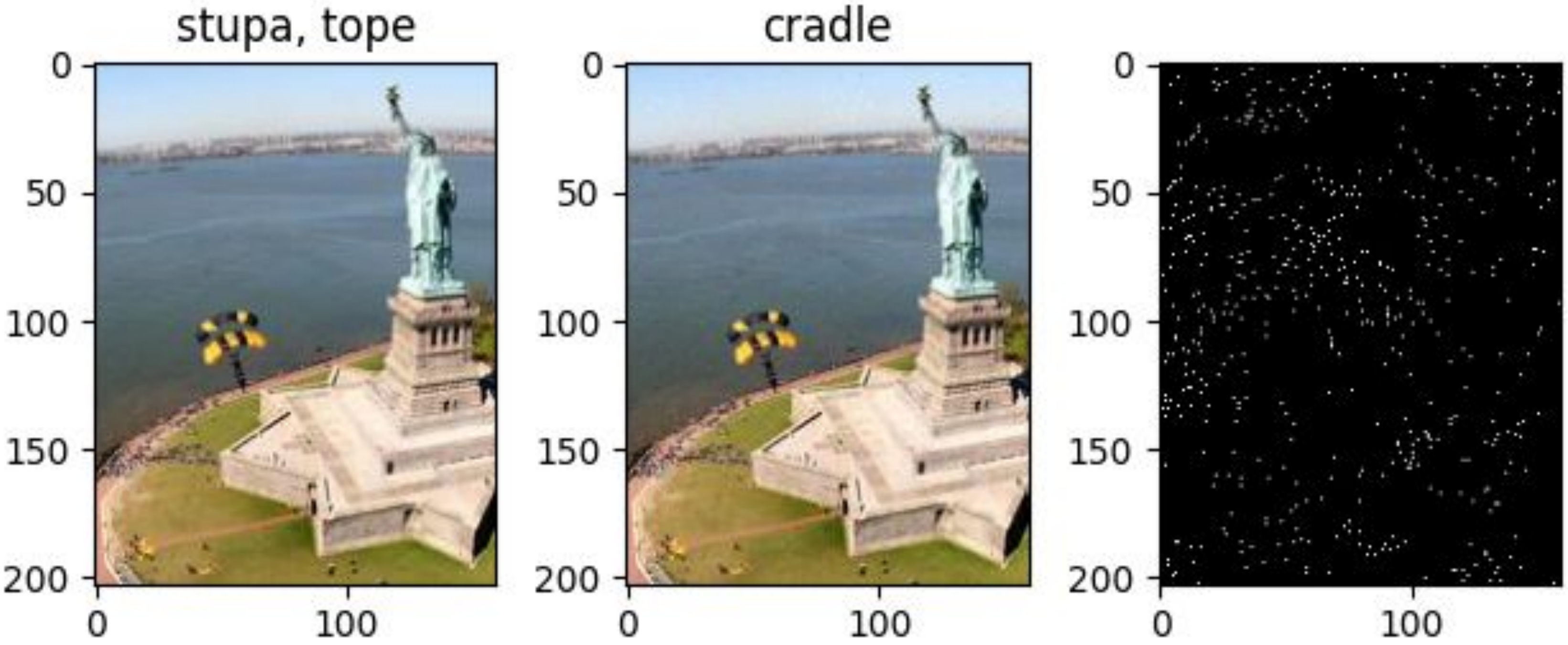
Visualization of pixel attack



Visualization of pixel attack



Visualization of pixel attack



Results on TinyImageNet (Untargeted Attack)

S.No	Epsilon	Attack Ratio	Avg. Iterations
1)	0.02	1.0	505
2)	0.1	1.0	126
3)	0.2	1.0	58
4)	0.3	1.0	39
5)	0.5	1.0	24

Results on TinyImageNet (Targeted Attack)

S.No	Epsilon	Attack Ratio	Avg. Iterations
1)	0.02	1.0	1331
2)	0.1	1.0	294
3)	0.2	1.0	162
4)	0.3	1.0	120
5)	0.5	1.0	79

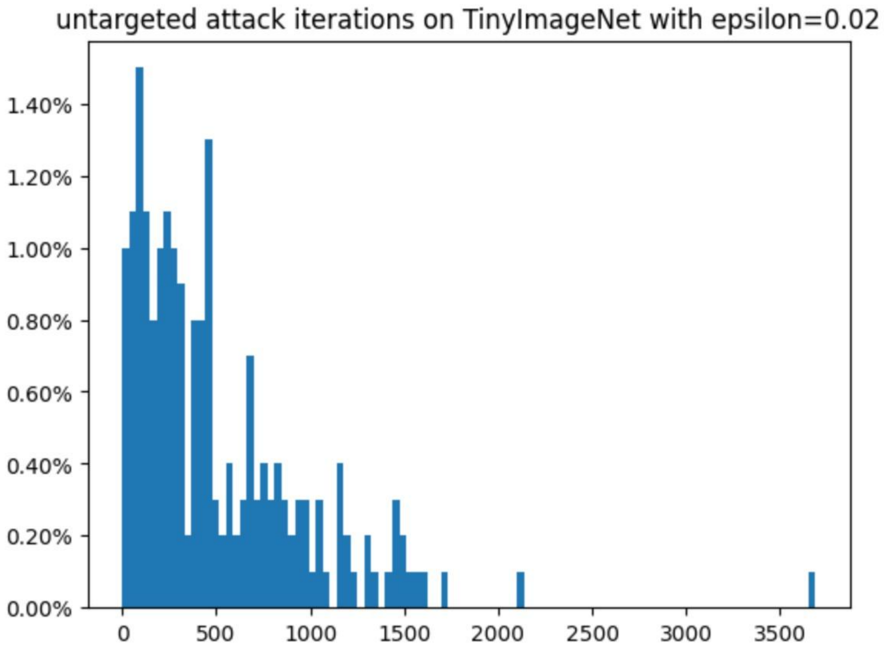
Results on ImageNet Subset (Untargeted Attack)

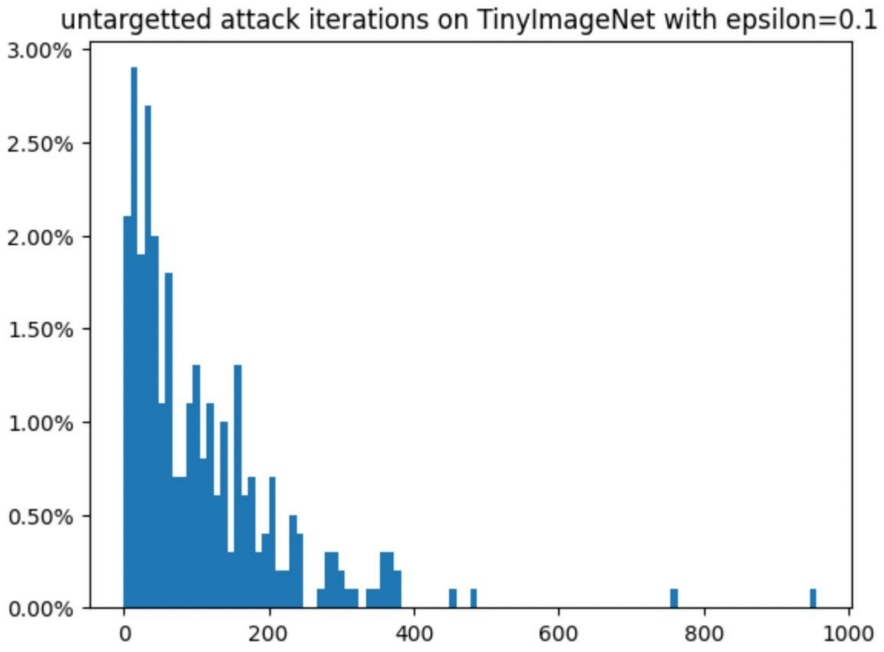
S.No	Epsilon	Attack Ratio	Avg. Iterations
1)	0.02	0.80	3126
2)	0.1	0.97	1648
3)	0.2	1.0	1240
4)	0.3	1.0	945
5)	0.5	1.0	658

Results on ImageNet Subset (Targeted Attack)

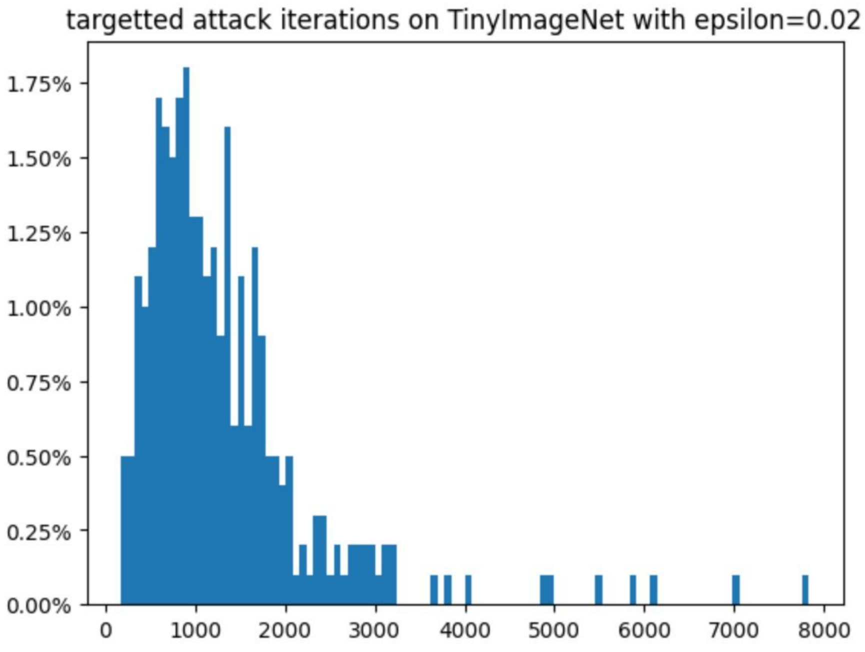
S.No	Epsilon	Attack Ratio	Avg. Iterations
1)	0.02	0.53	5383
2)	0.1	0.90	3383
3)	0.2	0.96	2824
4)	0.3	0.98	2636
5)	0.5	1.0	2410

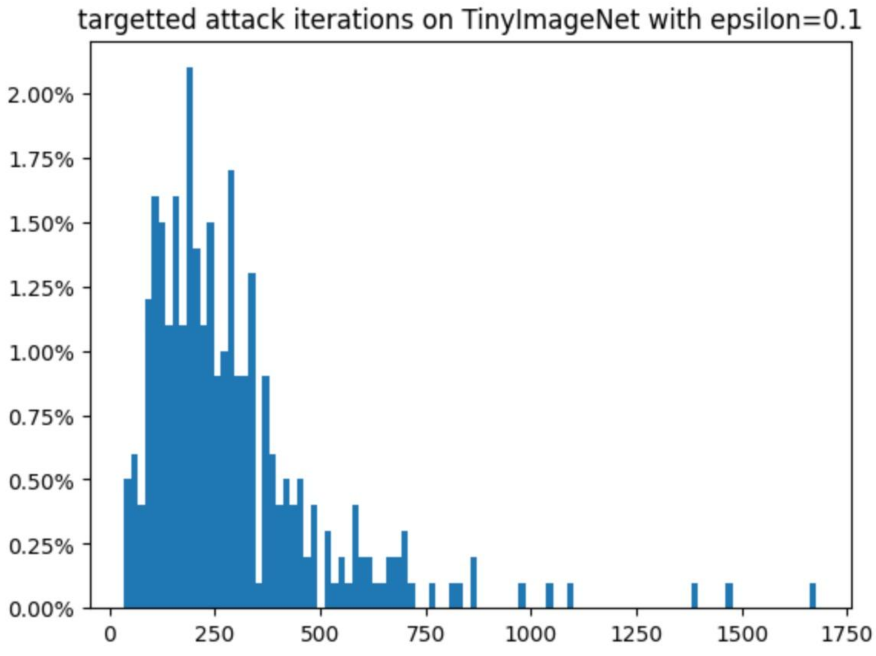
Distribution of untargeted attack iterations over images

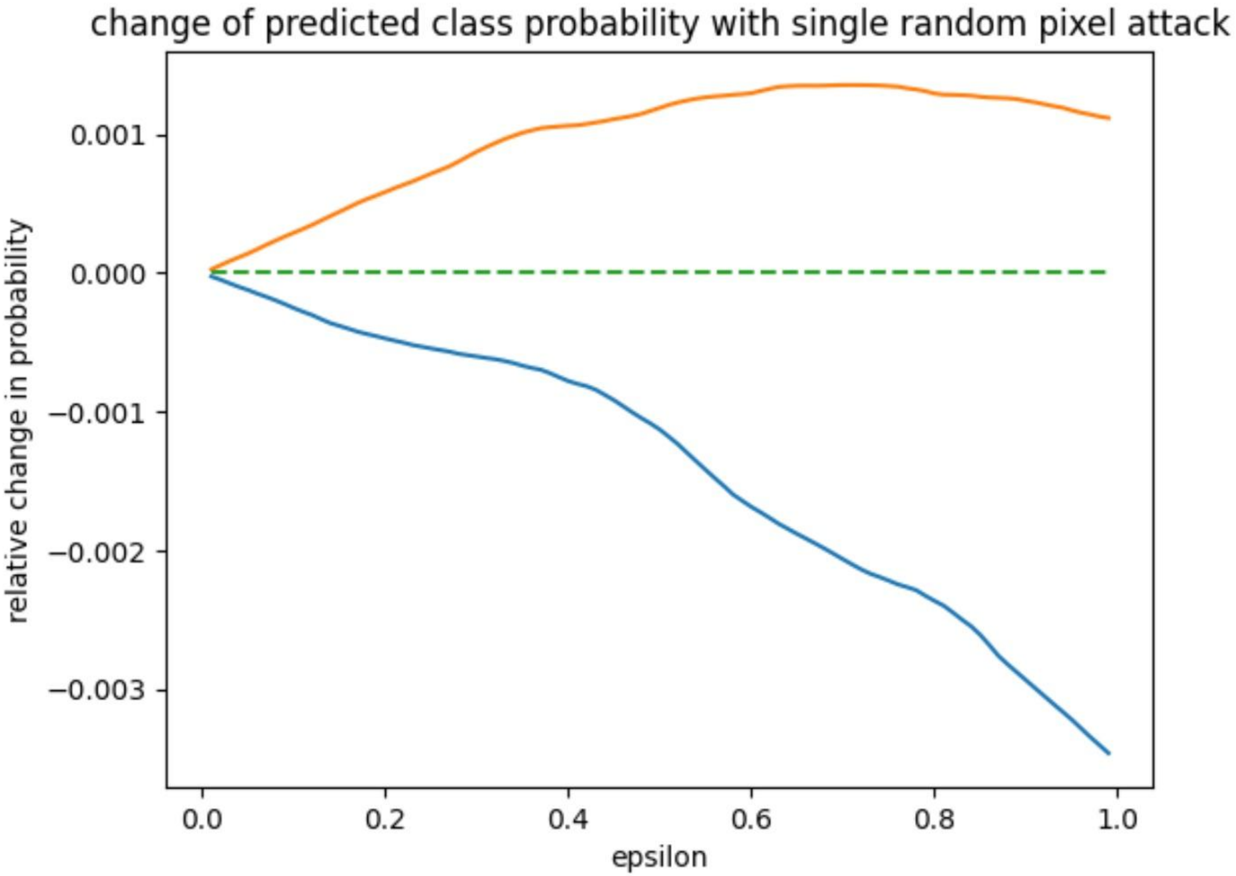


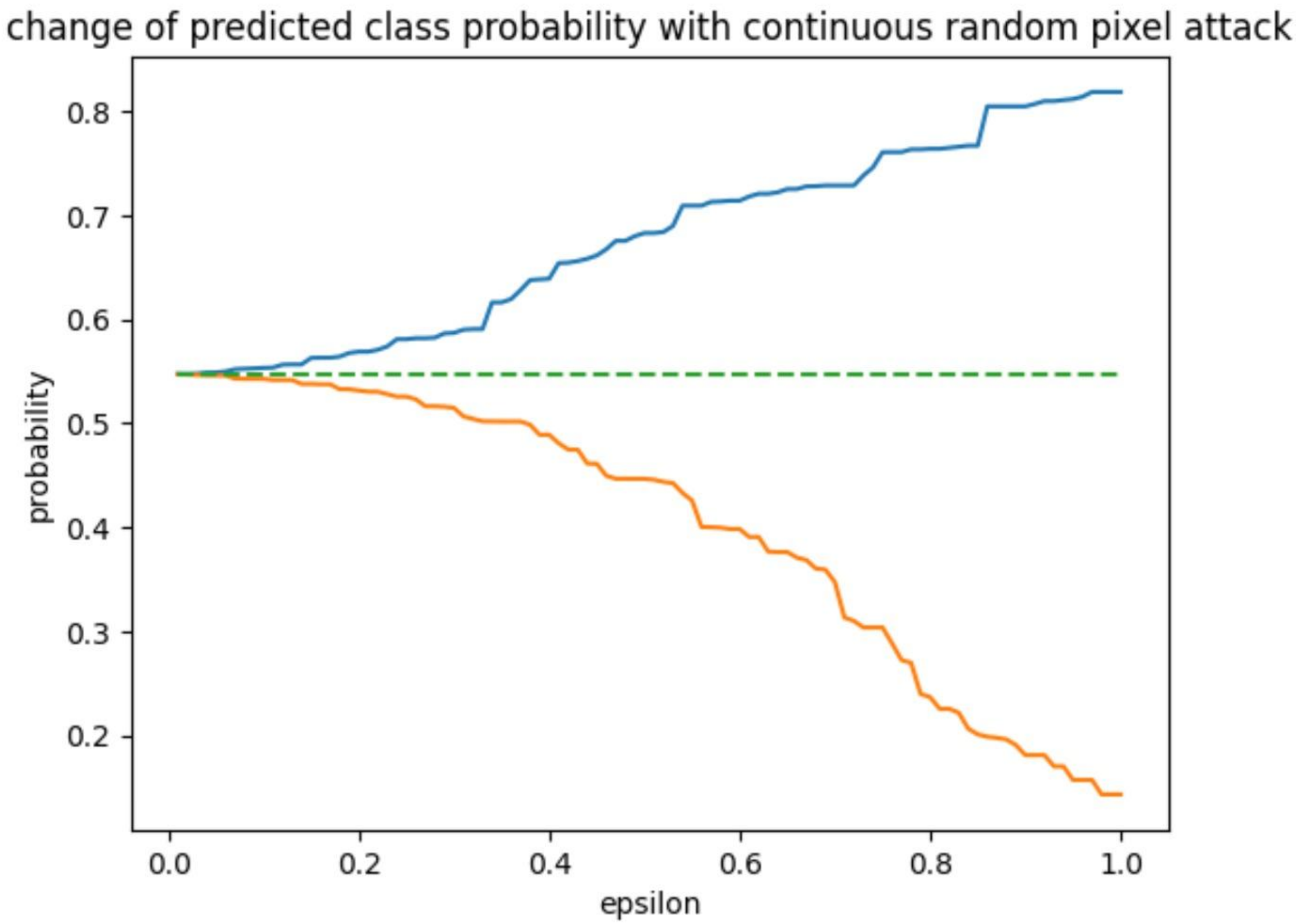


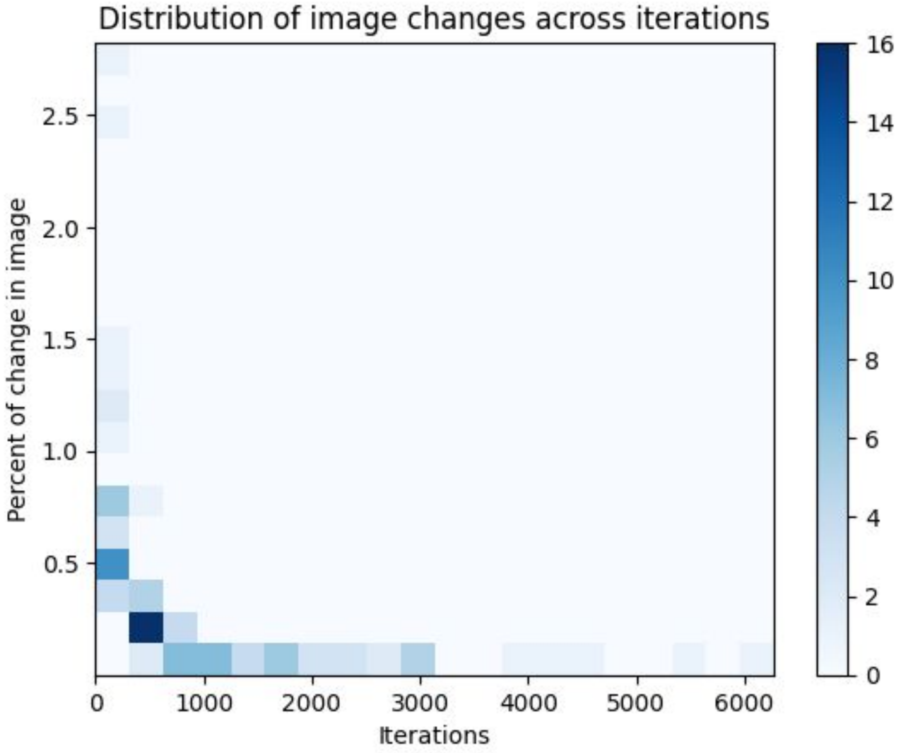
Distribution of targeted attack iterations over images



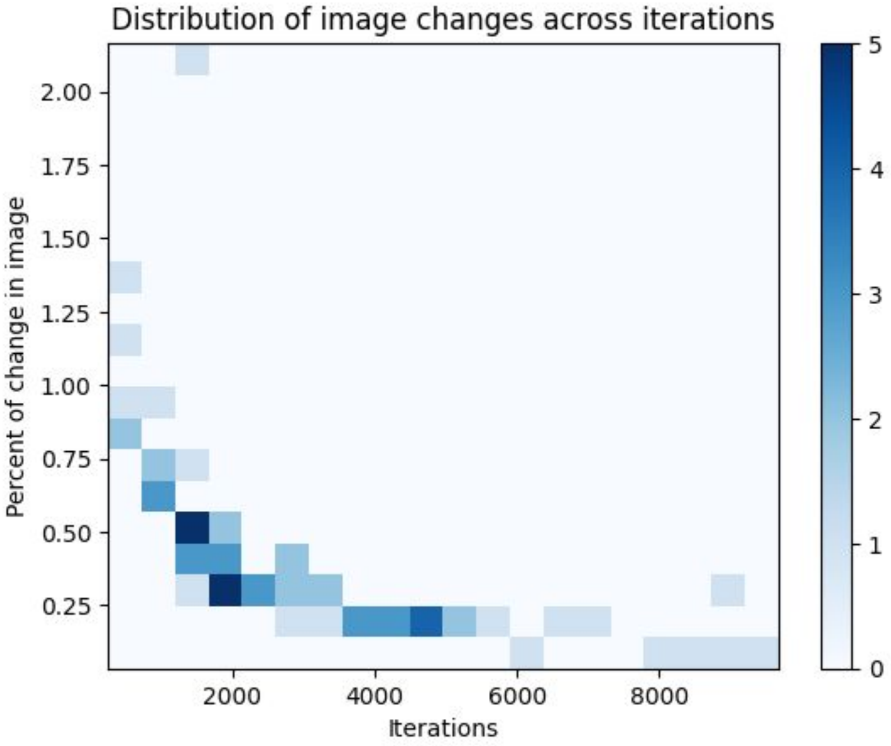



change of predicted class probability with single random pixel attack
relative change in probability
0.001
0.000
−0.001
−0.002
−0.003
0.0
0.2
0.4
0.6
0.8
1.0
epsilon


change of predicted class probability with continuous random pixel attack
0.8
0.7
0.6
0.5
0.4
0.3
0.2
probability
0.0
0.2
0.4
0.6
0.8
1.0
epsilon

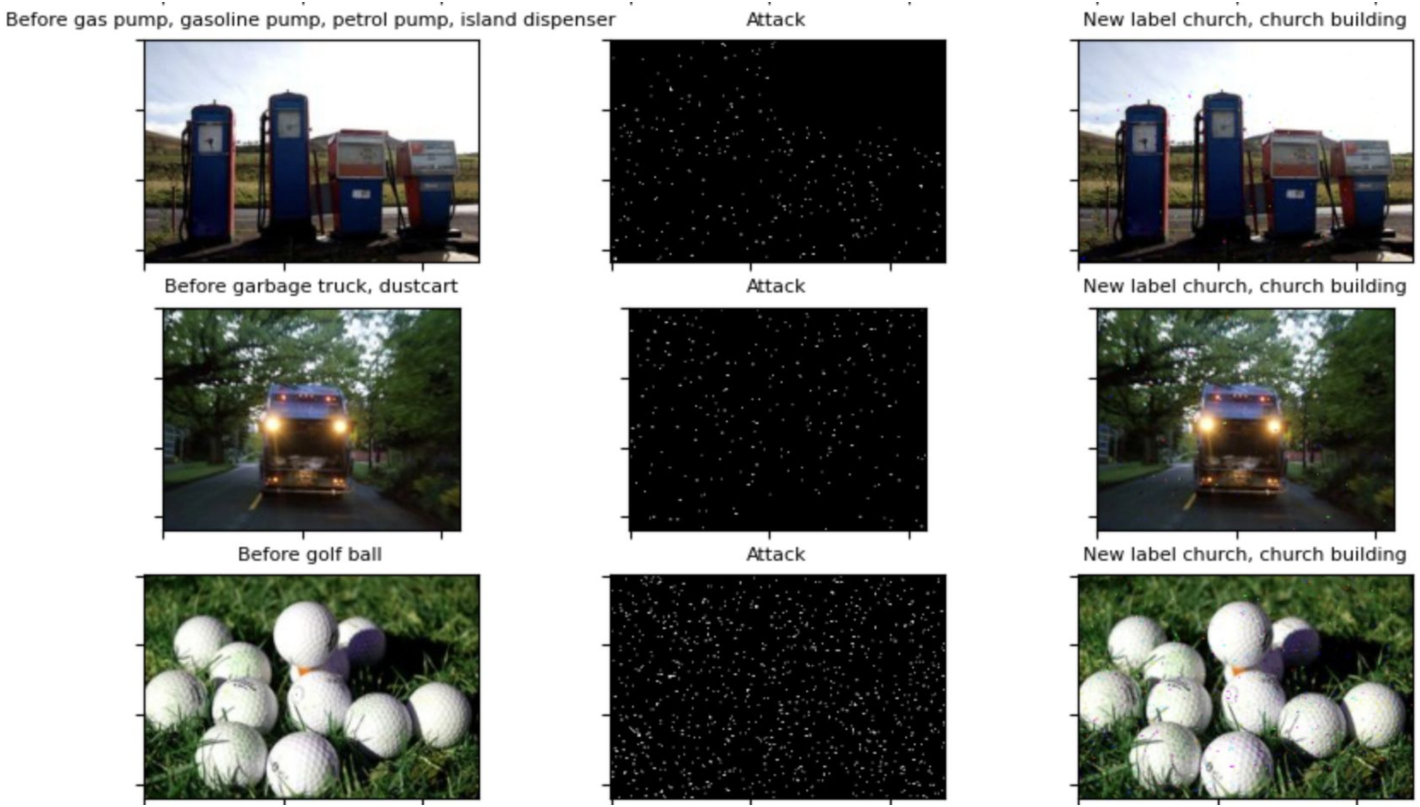


Untargeted attack, 100 images, eps=0.2



Targeted attack, 61 images, eps=0.2

Improvements in strategy ~ <1000 iterations



Contribution:

Ardhendu Banerjee: Prepared the code for Black box pixel attack, extracted data subsets and label mappings for experimentation; conducted initial tests on multiple pretrained models for evaluating attack resistance v/s query time tradeoffs

Animesh Das: Compared various strategies for attack (localization) and conducted targeted attack visualization and performance tests; result compilation

Ritvik Gupta: Prepared code for attack on TinyImageNet and automated efficient runs for random image indices for tabulation and computation of success rate

Thank You