Adversarial Attacks to fool Deep Neural Networks

Advanced Machine Learning UE15CS421

Reshma Bhat 01FB15ECS235
Nikitha Rao BS 01FB15ECS364
Rutha Prasad 01FB15ECS367

Types of Attacks

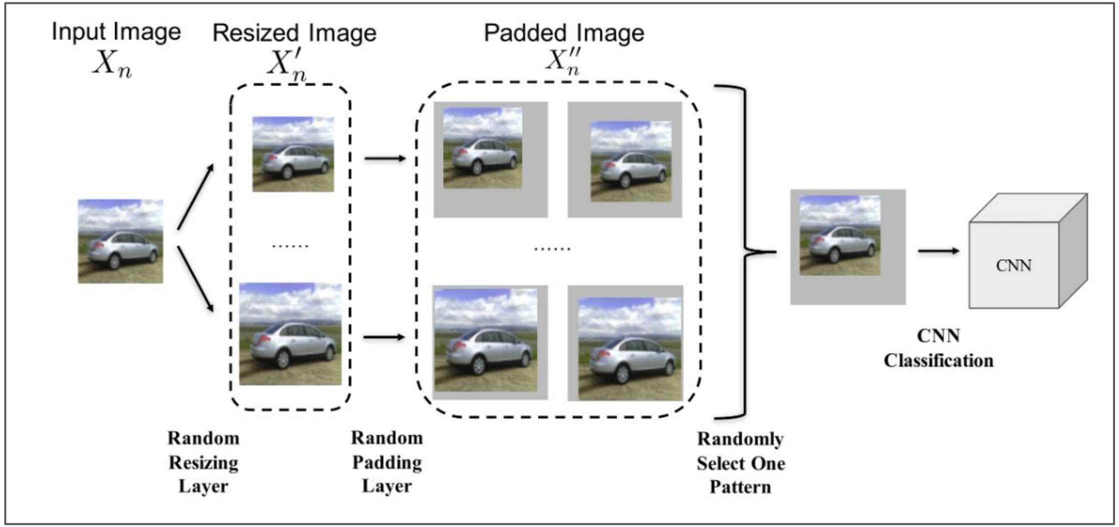
Non targeted Fast Gradient Sign Method

$$x^{adv} = x + \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{true}))$$

Defense Algorithm

GAN Discriminator

Random Resizing And Padding



Types of Attacks

Targeted Fast Gradient Sign Method

$$x^{adv} = x - \varepsilon \cdot \mathrm{sign}(\nabla_x J(x, y_{target}))$$

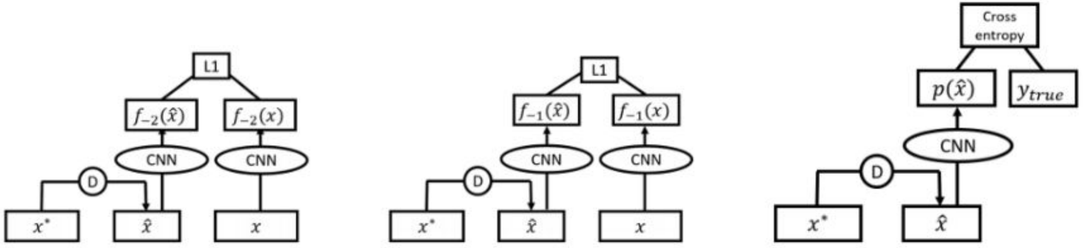
Defense Algorithm

Denoising features

- feature guided denoiser
- logits guided denoiser
- class label guided denoiser

Max Mean Discrepancy training

Capsule Networks



(a) FGD (b) LGD (c) CGD

Types of Attacks

Non targeted One Pixel Attack

Minimize(P(true)) Maximize(P(other))

Targeted One Pixel Attack

Maximize(P(target))

Defence Algorithm

Ensemble Adversarial training

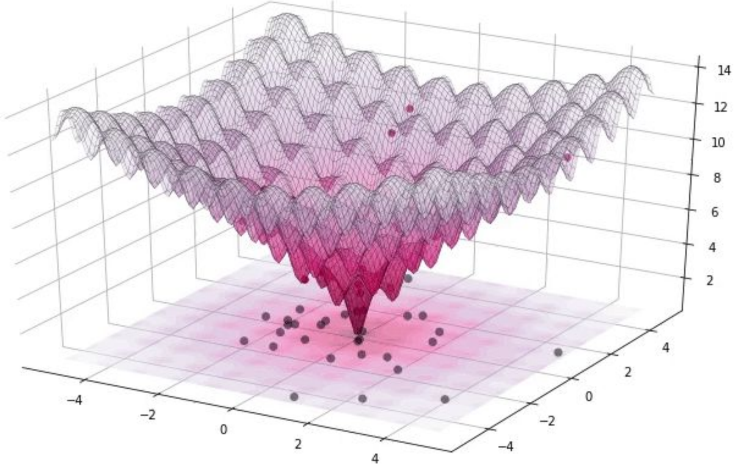
Momentum Matching

$$X = (x_1, y_1, r_1, g_1, b_1, x_2, y_2, r_2, g_2, b_2, \dots)$$

$$\mathrm{P} = (X_1, X_2, \dots, X_n)$$

$$X_i = X_{r1} + F(X_{r2} - X_{r3})$$

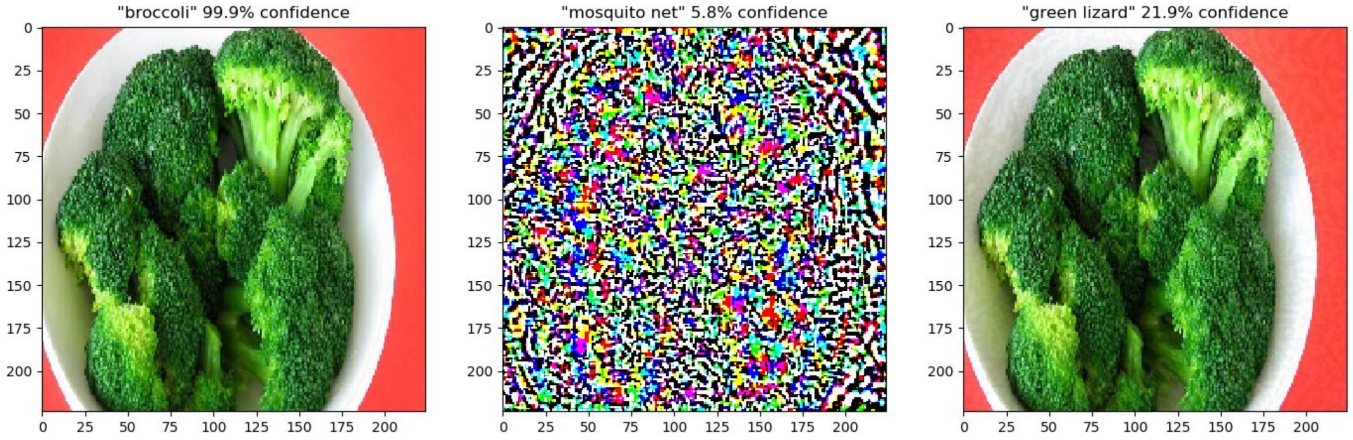
$$r1 \neq r2 \neq r3$$



Differential evolution on Ackley function

Results

FGSM

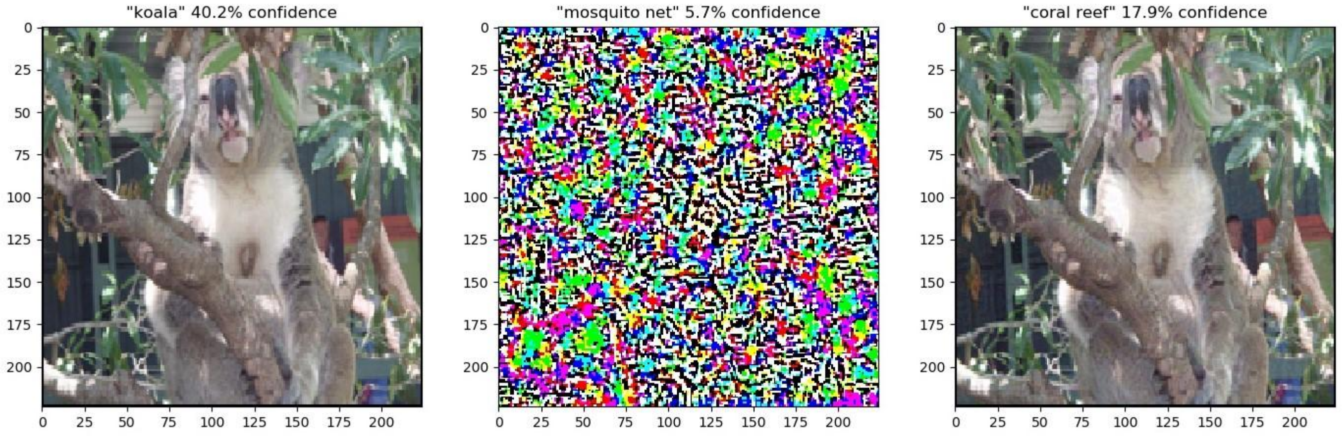


"broccoli" 99.9% confidence

"mosquito net" 5.8% confidence

"green lizard" 21.9% confidence

T-FGSM

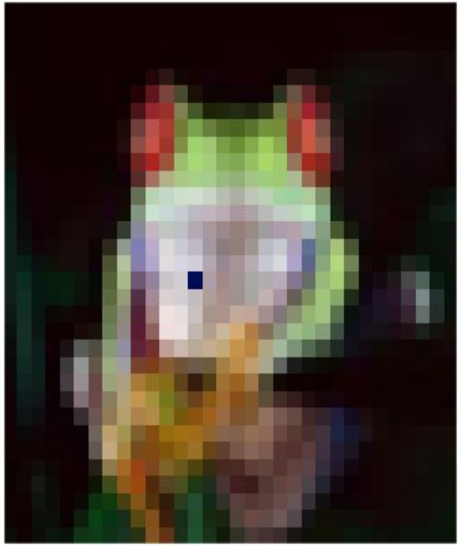


"koala" 40.2% confidence

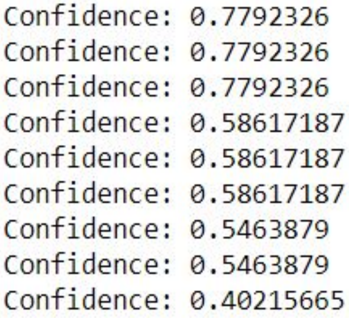
"mosquito net" 5.7% confidence

"coral reef" 17.9% confidence

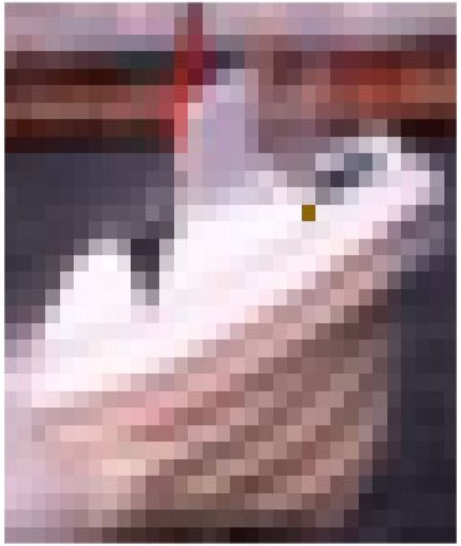
Results- One pixel attack and Targeted one pixel attack



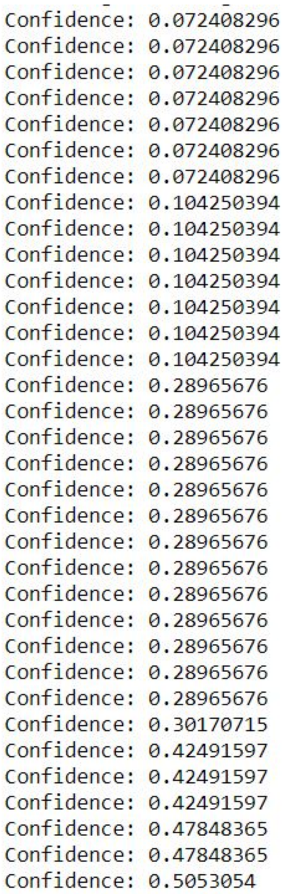
True: frog
Predicted: dog



```
Confidence: 0.7792326
Confidence: 0.7792326
Confidence: 0.7792326
Confidence: 0.58617187
Confidence: 0.58617187
Confidence: 0.58617187
Confidence: 0.5463879
Confidence: 0.5463879
Confidence: 0.40215665
```

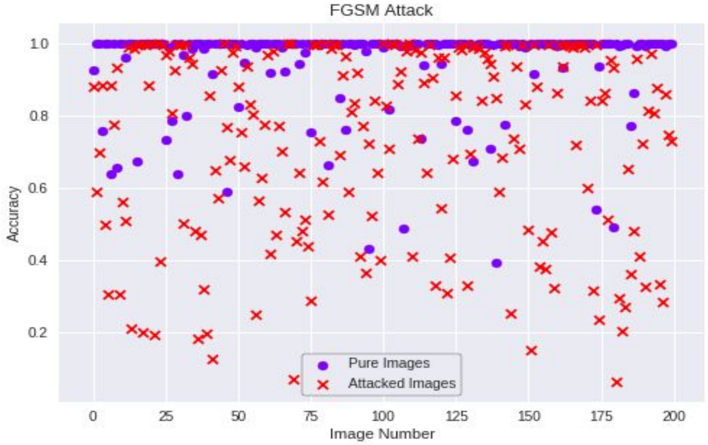


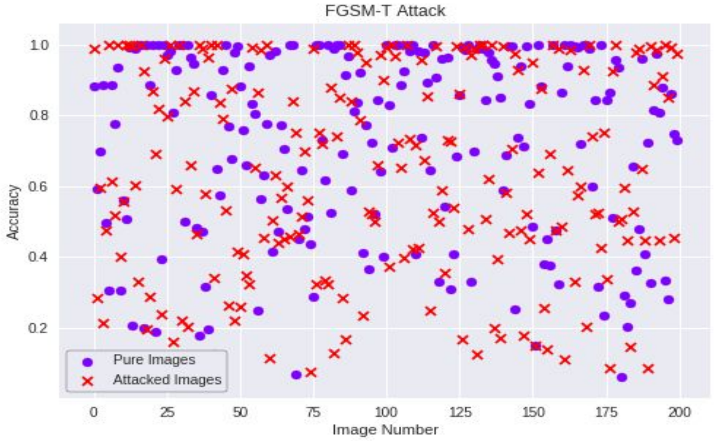
True: ship
Predicted: cat

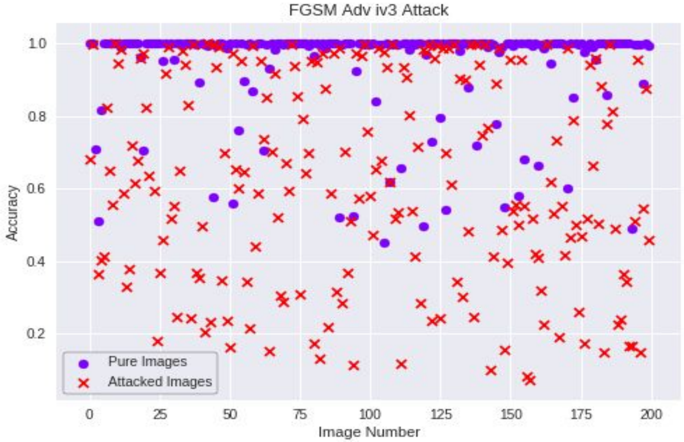


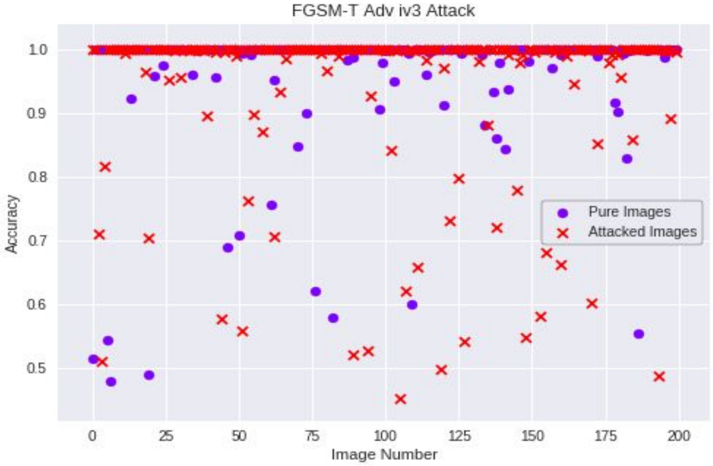
```
Confidence: 0.072408296
Confidence: 0.072408296
Confidence: 0.072408296
Confidence: 0.072408296
Confidence: 0.072408296
Confidence: 0.072408296
Confidence: 0.072408296
Confidence: 0.104250394
Confidence: 0.104250394
Confidence: 0.104250394
Confidence: 0.104250394
Confidence: 0.104250394
Confidence: 0.104250394
Confidence: 0.104250394
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.28965676
Confidence: 0.30170715
Confidence: 0.42491597
Confidence: 0.42491597
Confidence: 0.42491597
Confidence: 0.47848365
Confidence: 0.47848365
Confidence: 0.5053054
```

Comparison IncpV3 v/s Adv_IncpV3









Experiments and Results

- VGG16 - for feature extraction from pure and adversarial images
- Plot the extracted features to show contrast between the above two classes

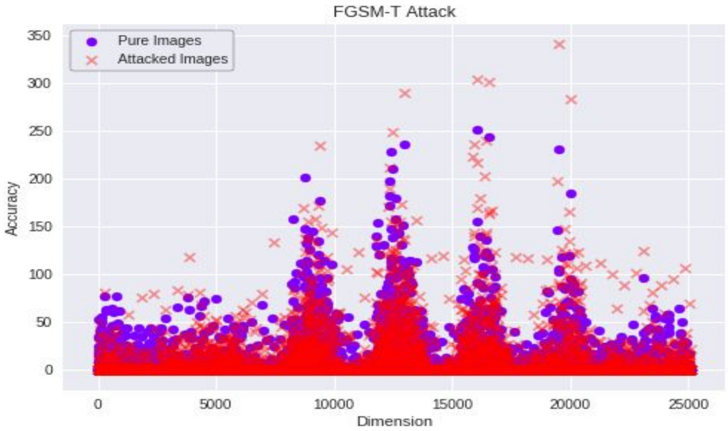


Fig 1: Contrast for FGSM-T Attack

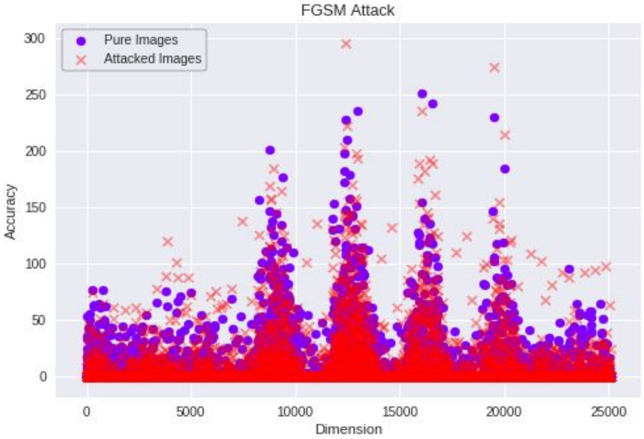


Fig 2: Contrast for FGSM Attack

Thank You!