WhiteBox Attacks on Image Classification Model

B05902121 黃冠博、B05902113 陳宏昇
B05902120 曾鈺婷、B05902045 宋哲寬

Outline

- Attack on Model ?
- FGSM Attack
  - 基本介紹
  - 改進方法

Outline

- One Pixel Attack
  - 基本介紹
  - 討論與防禦
- Adversarial Training Defense

Attack on Model

1

Problem Description

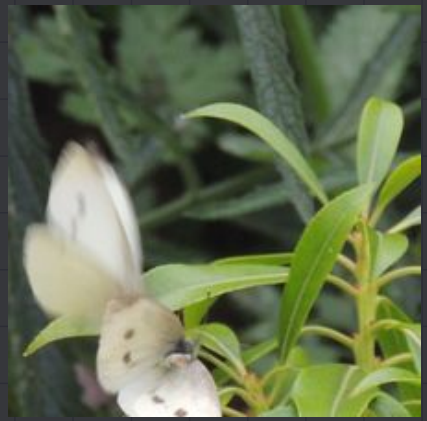
- 給定一個已知內部結構跟演算法的 whitebox model，針對這個 model 的 input 做一些改動，好讓原本能正確辨識圖片的 model 因為這些 adversarial image 出現預測錯誤的情況。

Threat Model

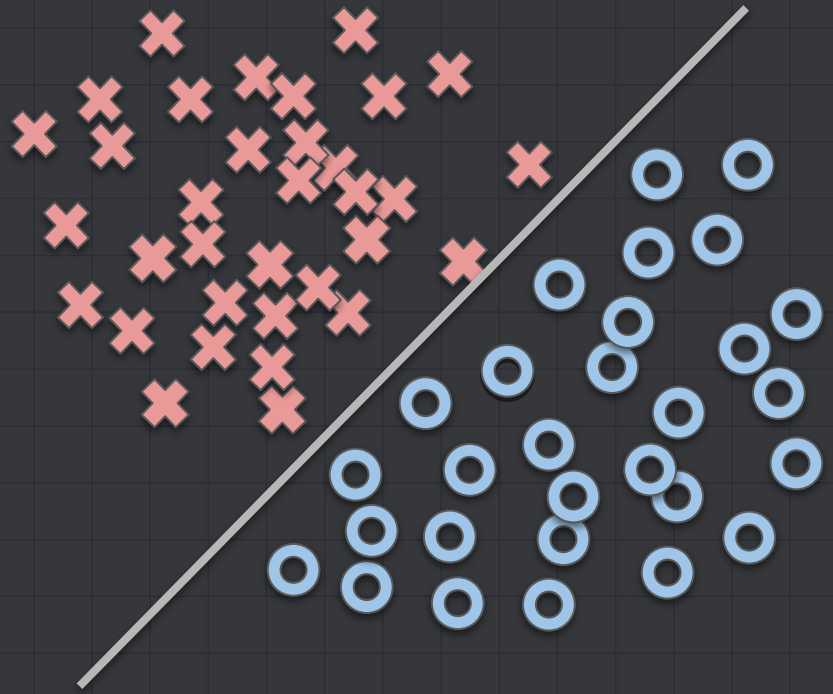
- **已知資訊**
  - 哪一種 model，如：resnet50, dense121, vgg19 ...
  - model 的 data preprocessing，如：圖片 normalize
- **能力範圍**
  - 持有一些能被 model 正確辨識的 input
  - 對 input 任意修改，並讓 model 做預測

Dataset

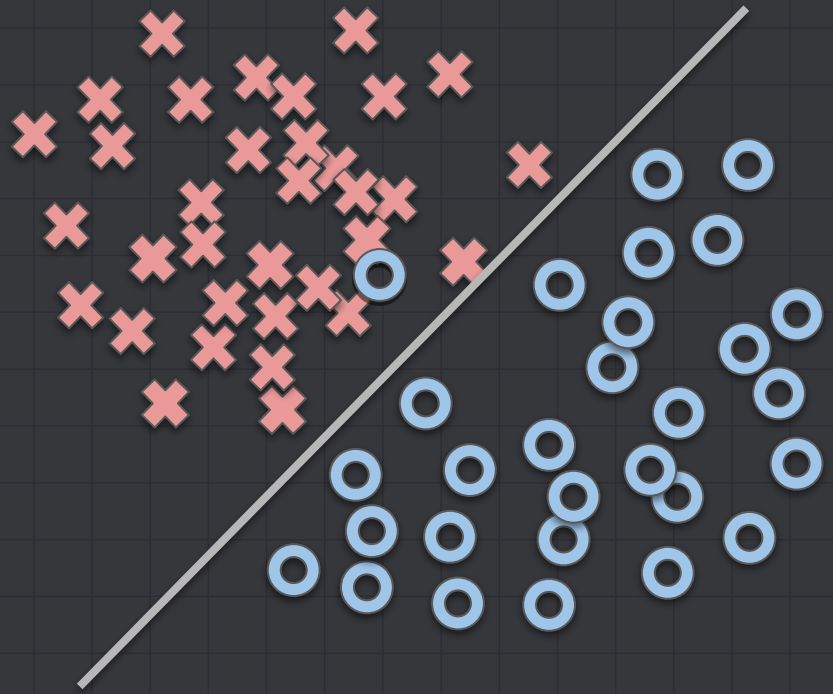
- MNIST (手寫數字)
- ImageNet 200 張圖
- ImageNet 若干張狗



Simple Example

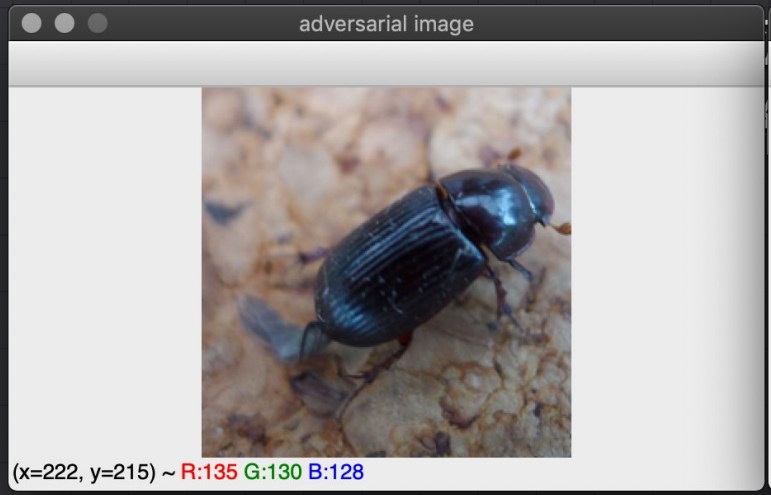


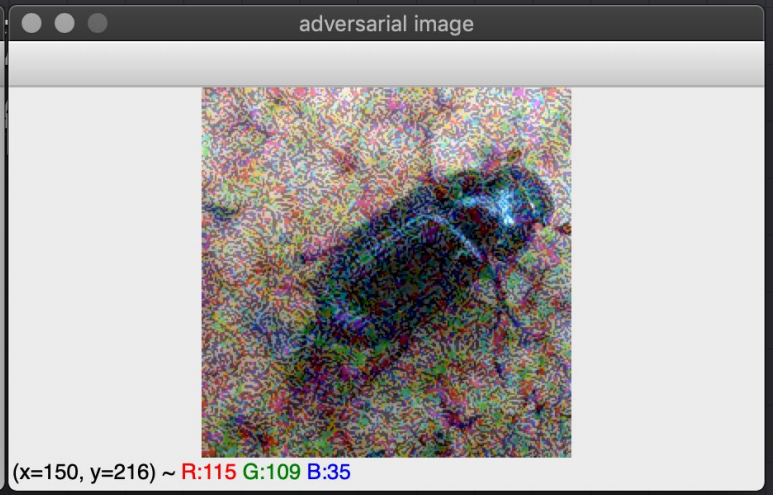
Simple Example



兩個衡量指標

- 成功率
  - 只要model 將某張圖片預測成不同於 ground truth





兩個衡量指標

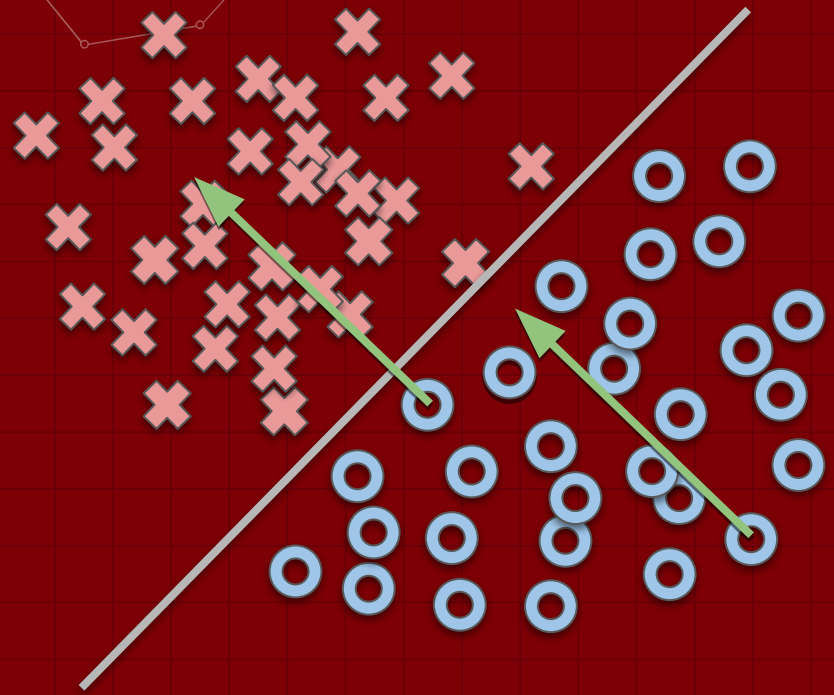
- 成功率
  - 只要model 將某張圖片預測成不同於 ground truth label 的類別就算攻擊成功。
  - 成功攻擊的圖片張數 / 總圖片張數

兩個衡量指標

- L-infinity norm
  - 攻擊後的圖片 (adversarial image) 與原圖的最大差距
  - adversarial image 是原圖加上 noise，將 adversarial image 跟原圖的每個 pixel 相減，取其中的最大值，就是 L-infinity norm

Fast Gradient Sign Method Attack

Original FGSM Attack

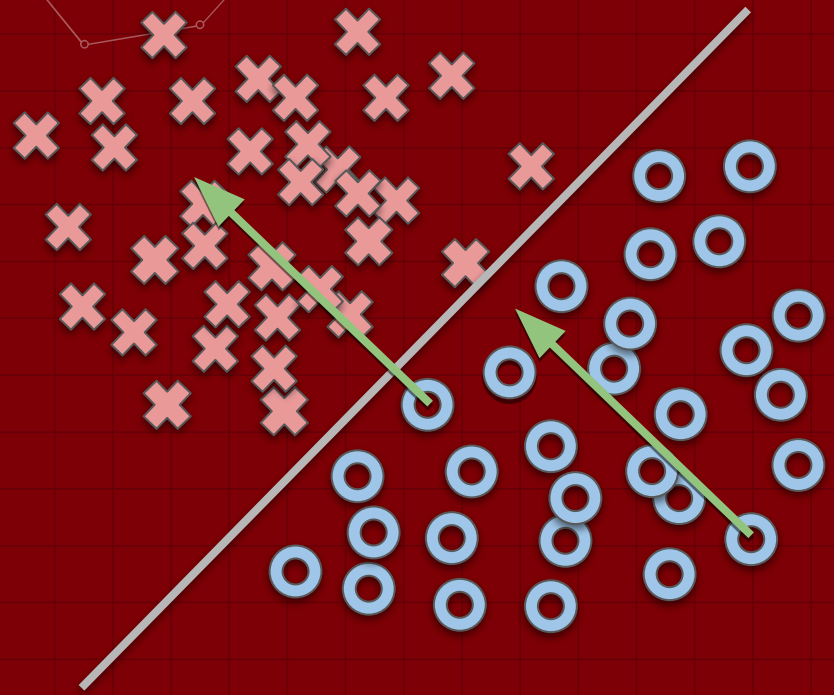


簡介

每個點想辦法往外走離開自己的類別

前進距離由 ε 決定

Original FGSM Attack



缺點

箭頭太短，無法跨越

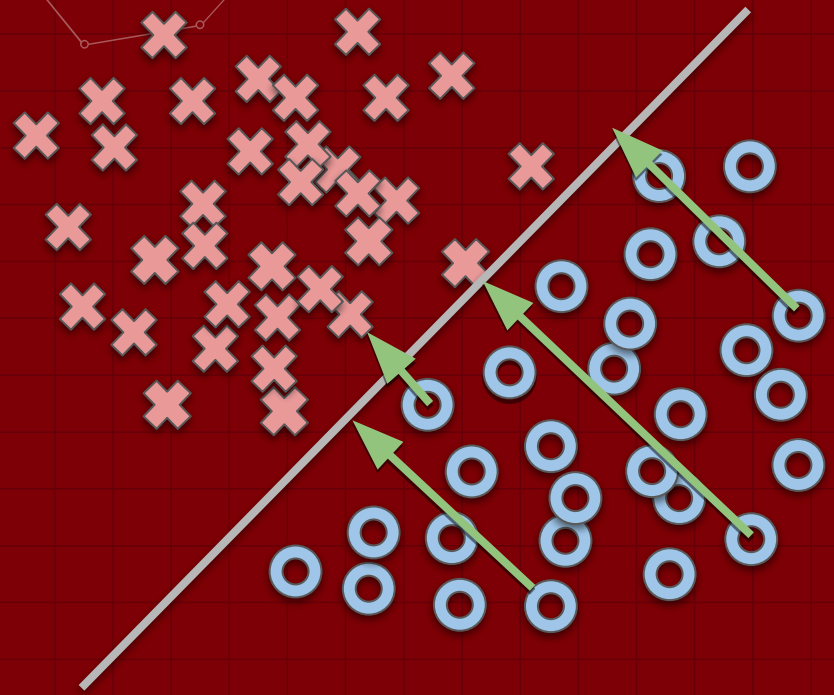
箭頭太長，攻擊前後差距太大 (L-inf 太大)

成功率

MNIST 0.990

ImageNet 0.985

Revised FGSM Attack

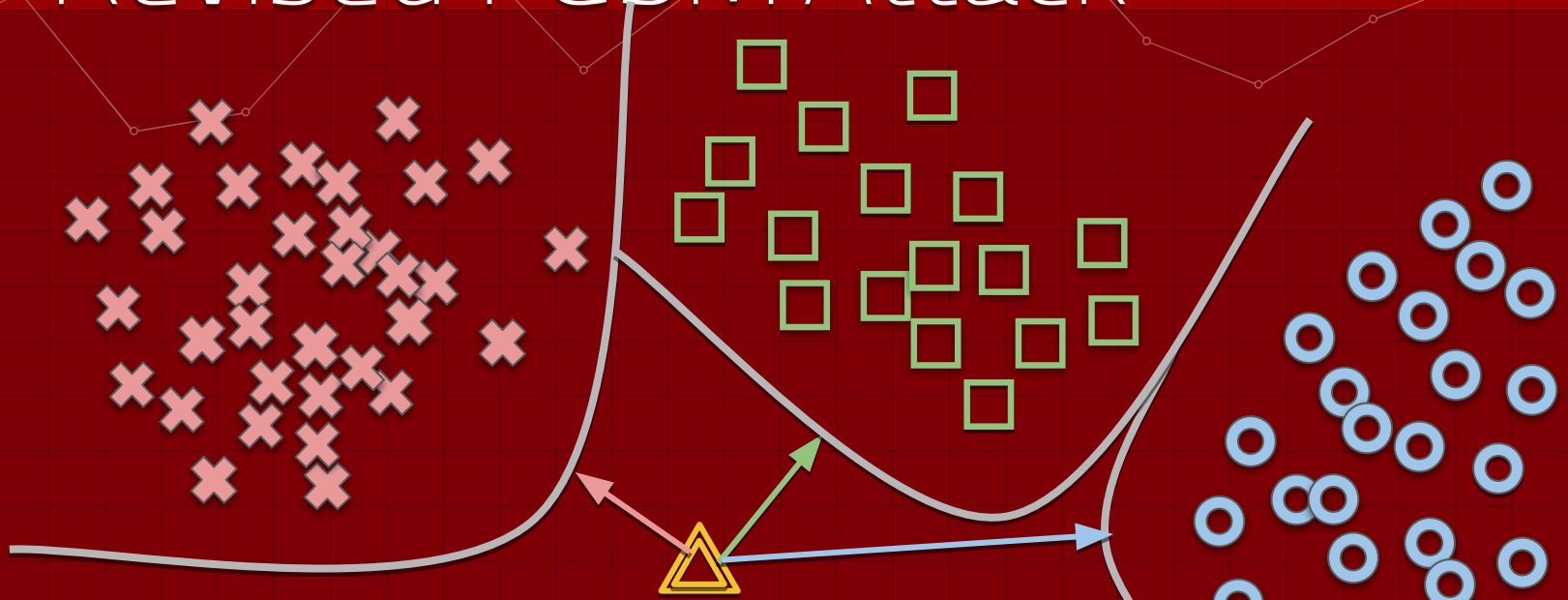


改進

每個點都給剛好跨越分割平面的 ε 就好

ε = 到分割平面的最短距離

Revised FGSM Attack



缺點 隨類別數上升，需要更多時間

成功率 MNIST 1.000、ImageNet 1.000

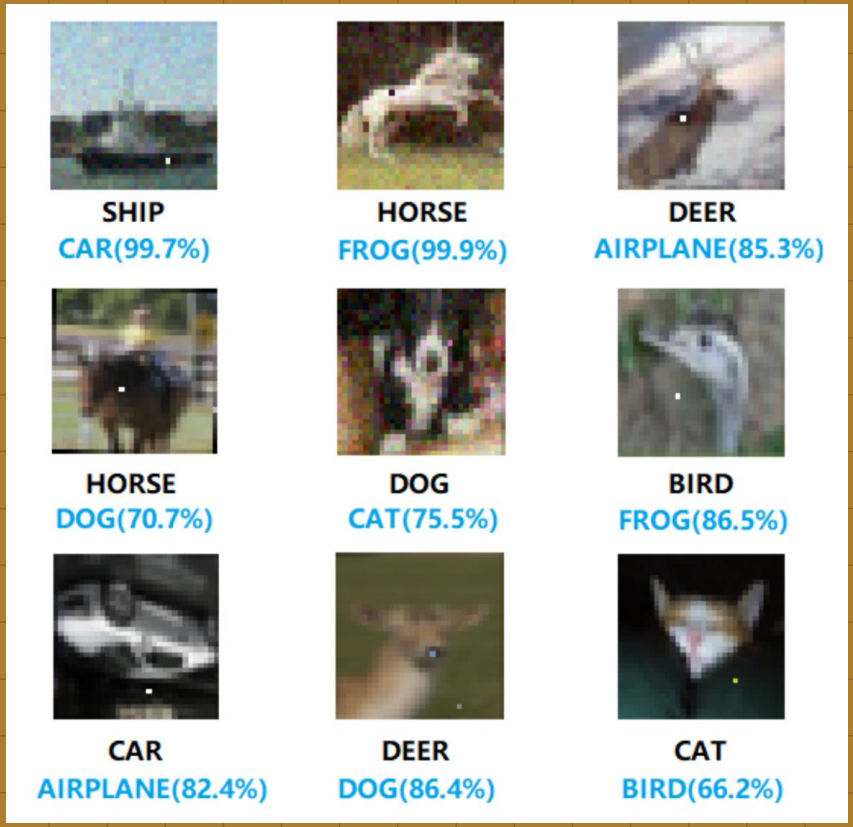
Results

resnet18

	成功率	L-inf
Original FGSM	MNIST 0.990 ImageNet 0.985	MNIST 5.00 ImageNet 5.00
FGSM + 改進方法	MNIST 1.000 ImageNet 1.000	MNIST 5.65 ImageNet 5.45

One Pixel Attack

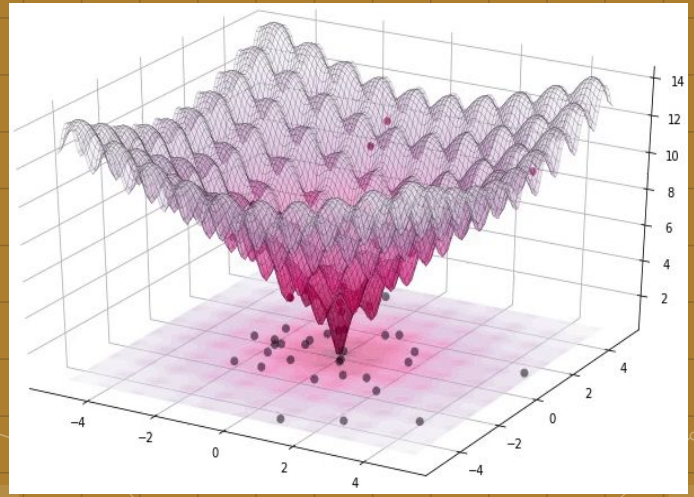
3



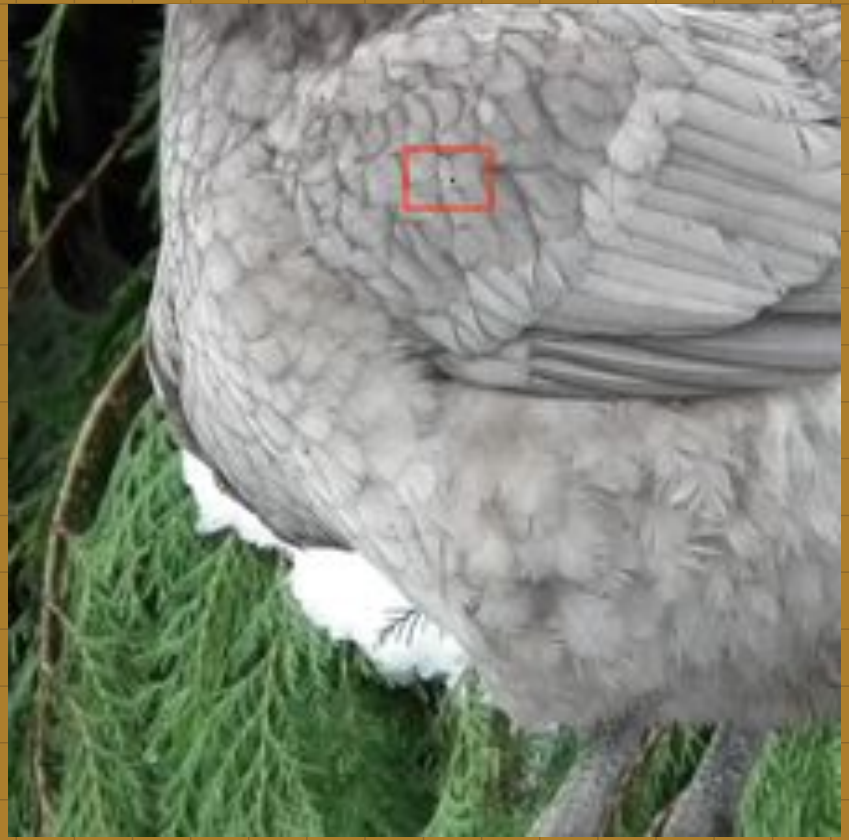
簡介 僅改變一個 pixel 而能夠使神經網路分類錯誤

Algorithm

- **遺傳演算法**
  - 多維度實數的最佳化問題
  - Differential Evolution
    - 非梯度
    - Global Maximum
    - Greedy
    - 突變、交叉、淘汰
  - 最大遞迴 iter、子代生成數 popsize



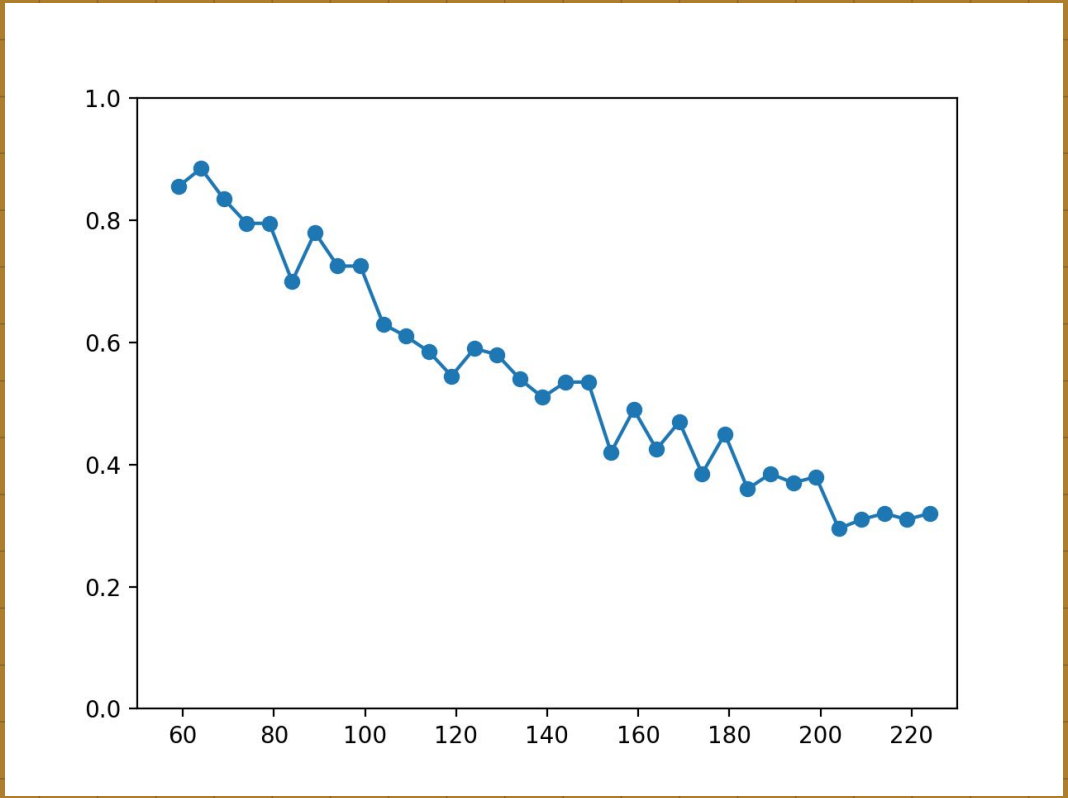
iter = 50; popsize = 100; 成功率 = 0.53



Discussion and Results

- **相關討論**
  - 成功率與圖片大小的關係





成功率與圖片大小的關係

Defense

- **防禦機制**
  - Filter
  - Random Change
  - Discontinuity

Defense- Filter

Defense- Filter

- Target 成功率
  - 把辨識結果變成某個class A
  - targeted def % = 0.74
- Untarget 成功率
  - 使辨識結果判斷錯誤
  - untargeted def % = 0.37

Defense- Random Change

- **簡介** 隨意更改某 pixel 的值
- num = 1
  - targeted def % = 0.14
  - untargeted def % = 0.14
- num = 5
  - targeted def % = 0.40
  - untargeted def % = 0.33
- num = 10
  - targeted def % = 0.48
  - untargeted def % = 0.375

Defense- Discontinuity

- 簡介　攻擊的 pixel 會有高度的不連續性
- Target 成功率
  - targeted def % = 1
- Untarget 成功率
  - untargeted def % = 1

Adversarial Training Defense

4

Preface

- **實驗動機**
  - model 可以被攻擊，代表有訓練不足之處
  - 利用 training 提升 model 的抵禦能力
- **實驗目標**
  - train 出更加 robust 的 model

Experiment

- 先搜集 image data
- 使用 pretrained model 標記資料
- FGSM 生成 adversarial data
- 將原本的 data 跟 adversarial data 拿回去 train

Data

- ImageNet 上搜集資料
- 使用 resnet18 為初始 model
- 生成 5500 張 adversarial data 下去 train
- 保留相似類別的 500 張圖片做為 test

Results

- 對於原圖的防禦性
  - 使用原圖對新的 model 進行 FGSM
  - 對照組 resnet18
    原圖 FGSM 成功率 0.985
  - 實驗組 adversarial trained model
    原圖 FGSM 成功率 0.47

Results

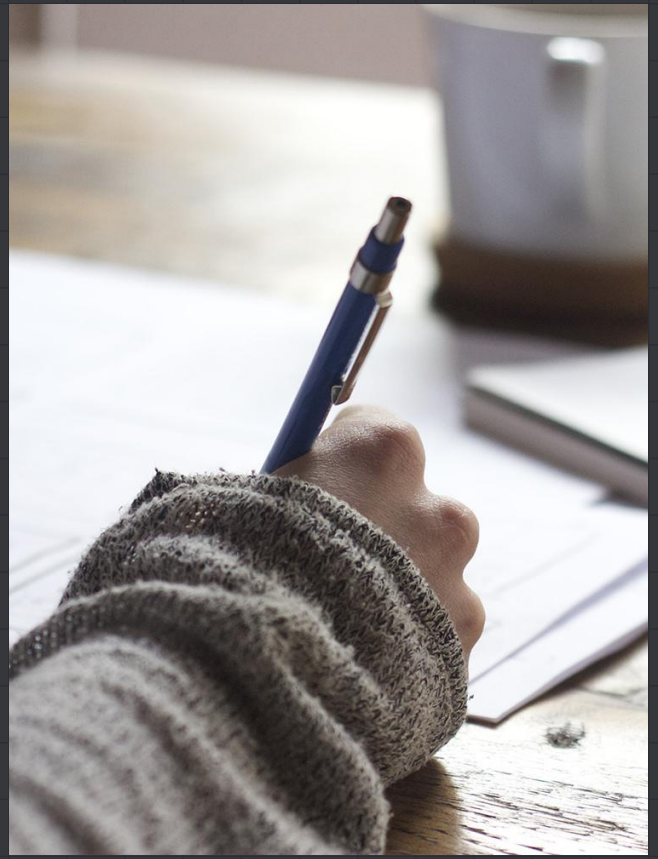
- 對於未知資料的防禦性
  - 同一 label 資料對於原圖進行 FGSM
  - 對照組 resnet18
    保留資料 FGSM 成功率 0.985
  - 實驗組 adversarial trained model
    保留資料 FGSM 成功率 0.81

Conclusion

- FGSM 與 revised FGSM，是成功率及 L-inf 做 tradeoff
- 明白圖片尺寸對於 one pixel attack 的影響，除此之外，也找到非傳統filter的有效防禦方法。
- 使用 adversarial training 後，有助於提升 model 強度
  是否有機會經過多次的 training 而收斂？

THANKS!

Any Questions ?

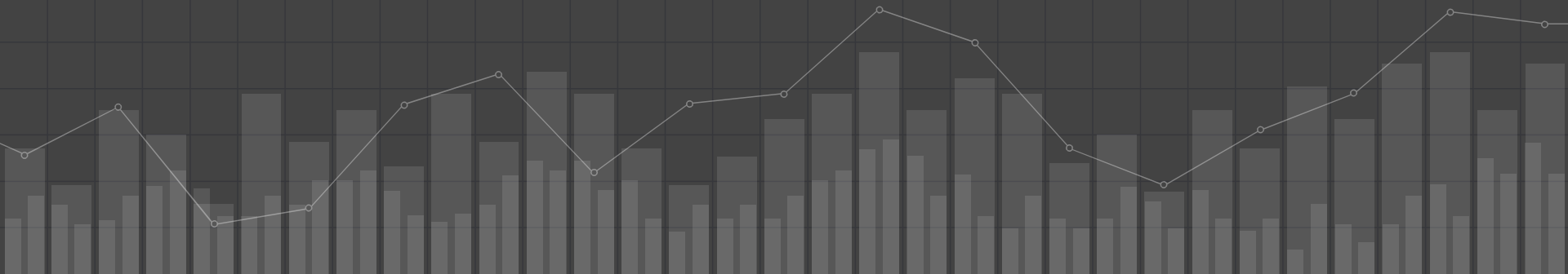


Appendix

5

Contribution

- Slides 所有人
- Fast Gradient Sign Method Attack 黃冠博
- Demo 曾鈺婷
- One Pixel Attack 陳宏昇
- Adversarial Training Defense 宋哲寬



Reference

- DeepFool: a simple and accurate method to fool deep neural networks Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard Ecole Polytechnique F´ed´erale de Lausanne
- One pixel attack for fooling deep neural networks,Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi, 2017