2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 20 – FGSM Attack

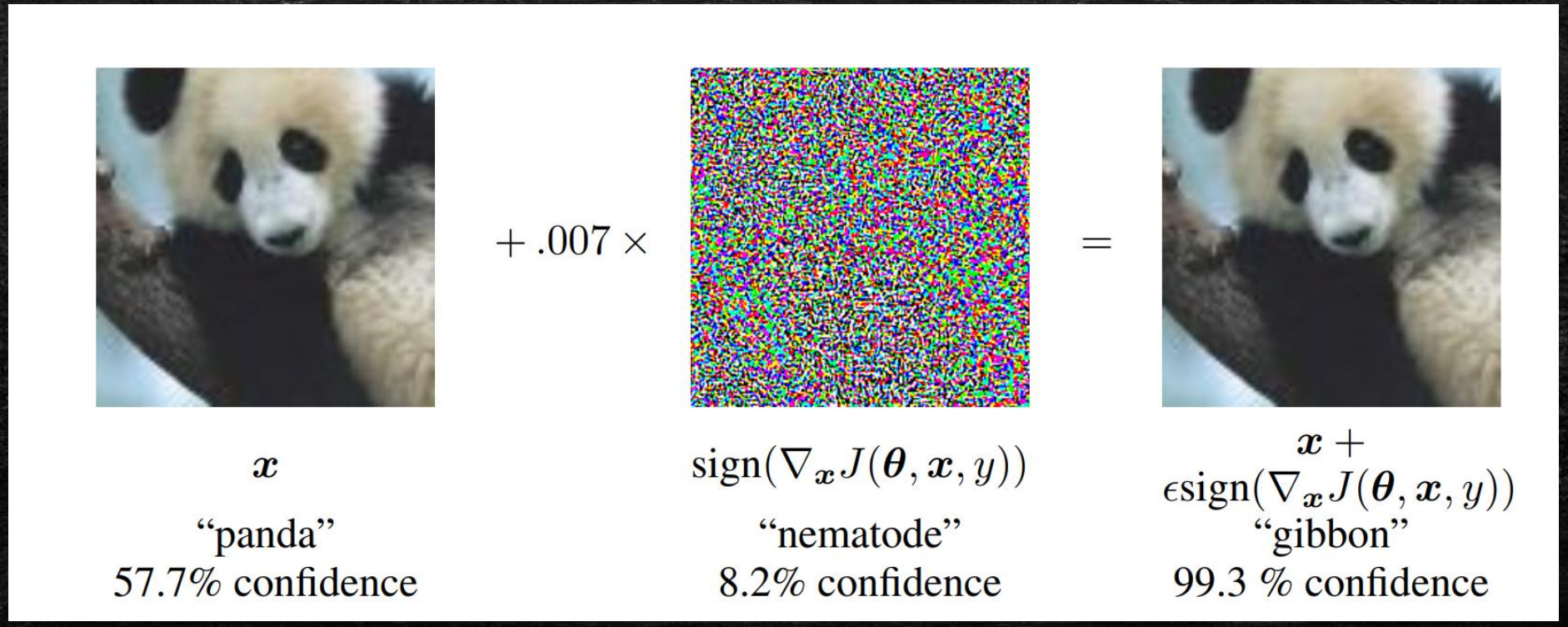
大綱

- FGSM 攻擊
  - 前情提要
  - 程式實作
- 結論

前情提要

https://arxiv.org/pdf/1412.6572

- EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES (2015)



前情提要

- 有趣的地方要來了，如果換個對象再來一遍呢

Let $\boldsymbol{\theta}$ be the parameters of a model, $\boldsymbol{x}$ the input to the model, y the targets associated with $\boldsymbol{x}$ (for machine learning tasks that have targets) and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ be the cost used to train the neural network. We can linearize the cost function around the current value of $\boldsymbol{\theta}$, obtaining an optimal max-norm constrained pertubation of

$$\boldsymbol{\eta} = \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right).$$

- 紅色框的部分代表當代入當前在 loss function 帶入當前模型參數、輸入圖片x、輸入圖片標籤y，對於輸入圖片 x 的 gradient
- 用白話文來說就是得到改動圖片對 loss function 的變化量，然後往該方向去調動輸入圖片，就能夠產生 FGSM 的攻擊樣本

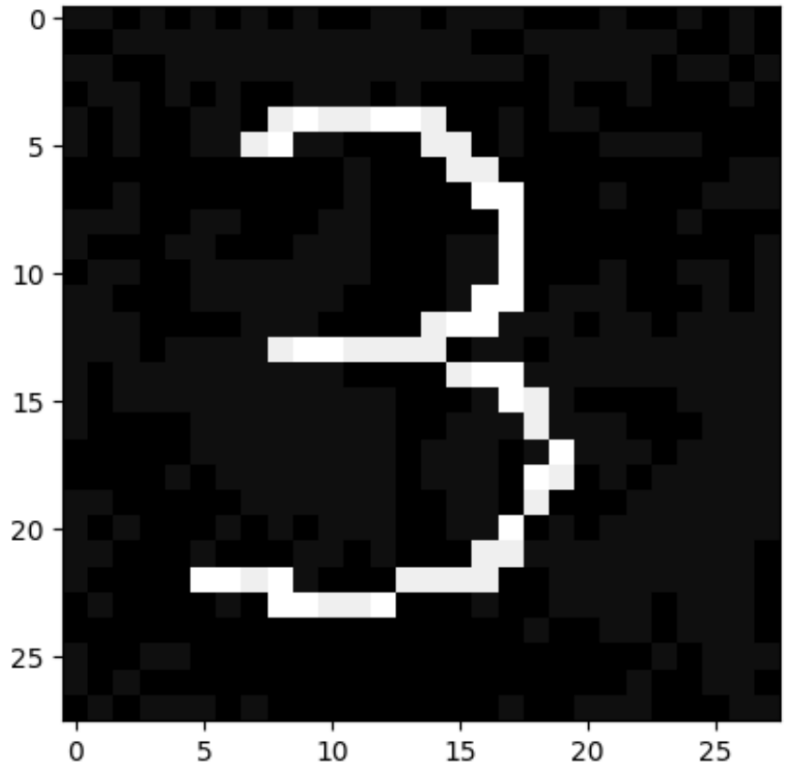
程式實作

https://www.tensorflow.org/tutorials/generative/adversarial_fgsm?hl=zh-cn

https://medium.com/berkeleyischool/fgsm-attacks-on-mnist-fashion-dataset-90cd0eeed7ab

```
1/1 [==============================] - 0s 12ms/step
model result: [[6.4199612e-06 4.8149347e-01 2.2625555e-01 2.8846627e-01 1.8393863e-06
  3.5566404e-03 1.1158231e-04 1.3270702e-05 9.2570641e-05 2.3361558e-06]]
predict value: 1

<matplotlib.image.AxesImage at 0x25634186c40>
```



結論

- FGSM 攻擊演算法的優點在於想法簡單好實作，但缺點在於無法指定要攻擊的目標，而且只是單純用梯度移動一次攻擊能力較弱

- 接下來打鐵趁熱，先稍微繞過對抗式攻擊樣本的範疇，先來針對梯度介紹模型梯度的資料洩漏演算法