2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 19 – FGSM Attack

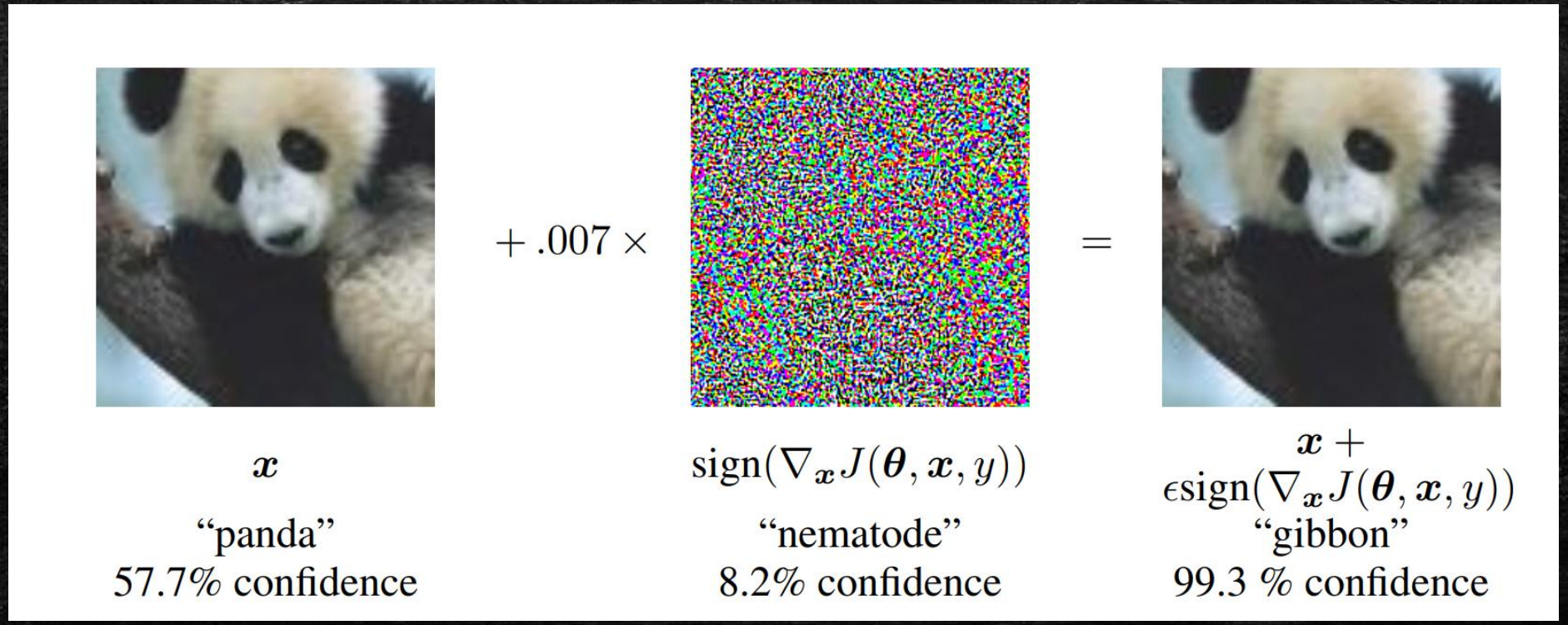
大綱

- FGSM 攻擊
  - 攻擊手法原理
  - 用程式算微分結果
  - 程式實作
- 結論

FGSM 攻擊手法原理

https://arxiv.org/pdf/1412.6572

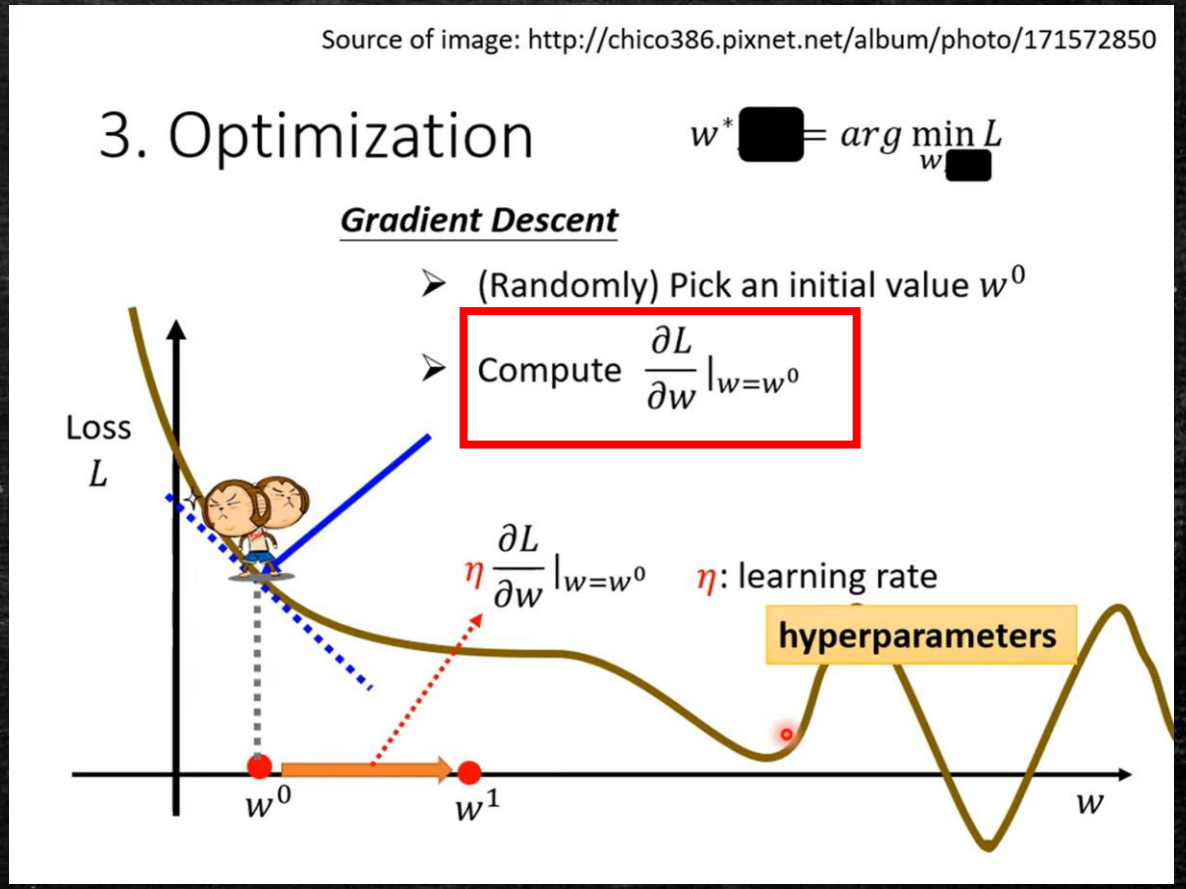
- EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES (2015)



FGSM 攻擊手法原理

- 我自己又稱它為叛逆演算法

- 稍微複習一下這個式子
  - 紅色框代表改動模型參數 w 對 Loss function 的變化量
  - 往它的反方向走去修改 w 會讓 Loss function 的數值變小



FGSM 攻擊手法原理

- 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

用程式算微分結果

https://www.tensorflow.org/guide/autodiff

- 推薦參考官方的教學資料，還蠻詳細的

```
x = tf.Variable(3.0)

with tf.GradientTape() as tape:
  y = x**2
```

```
# dy = 2x * dx
dy_dx = tape.gradient(y, x)
dy_dx.numpy()
```

```
6.0
```

當然也有更複雜的情況

- 遇到常數資料的，記得讓它被 watch 一下

```
x = tf.constant(3.0)
with tf.GradientTape() as tape:
  tape.watch(x)
  y = x**2

# dy = 2x * dx
dy_dx = tape.gradient(y, x)
print(dy_dx.numpy())
```

```
6.0
```

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

In [ ]:
```
# 開啟萃取圖檔稍微看一下樣子
img = Image.open("test.bmp")
test_data = np.asarray(img)
plt.imshow(test_data,cmap='gray')
img.close()

hack_img = test_data
```

In [ ]:
```
# 載入原本的模組
load_model = tf.keras.models.load_model('.\\mnist_basic_model.h5')
load_model.summary()
```

In [ ]:
```
# 想辦法透過 FGSM 演算法針對載入的圖檔做出生成式對抗樣本
```

In [ ]:
```
hack_img = hack_img.astype(np.uint8)
im = Image.fromarray(hack_img)
im.save("hack.bmp")

test_data = np.asarray(hack_img)
plt.imshow(test_data,cmap='gray')
```

結論

- FGSM 算是生成式攻擊樣本內數學最單純(?)的一個攻擊演算法
- 在實作數學類的演算法時務必了解所有符號的定義，要不然有時候差一點點就會讓整個演算法無法運作
- 透過這個演算法可以順便讓自己練習一下用 tensorflow 求導數的能力