2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 28 – Clean Label Attack

大綱

- Clean Label 攻擊
  - 前情提要
  - 程式實作

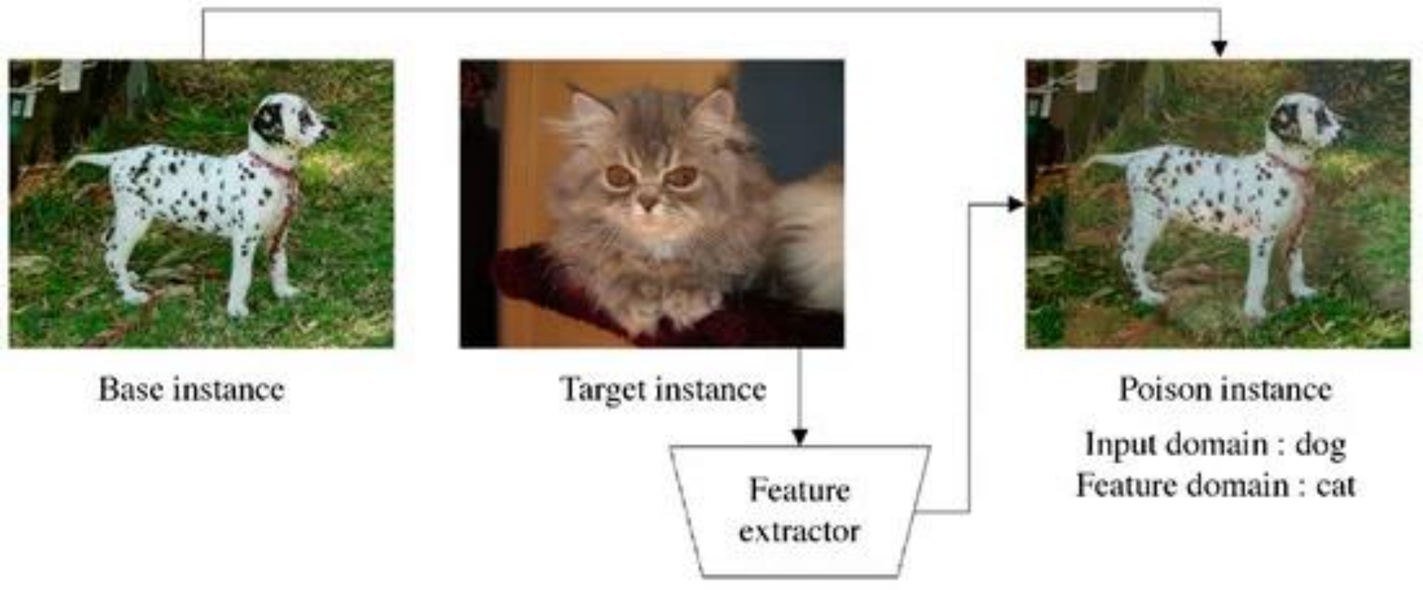
- 結論

長的跟原圖很像
$\beta \|\mathbf{x}-\mathbf{b}\|_2^2$
特徵跟目標很像
$\|f(\mathbf{x})-f(\mathbf{t})\|_2^2$
Clean Label 演算法
JAKE-CLARK.TUMBLR

前情提要

https://arxiv.org/abs/1804.00792

- Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks (2018)
- 簡單來說就是外表看似小孩，內在卻過於常人的名偵探樣本

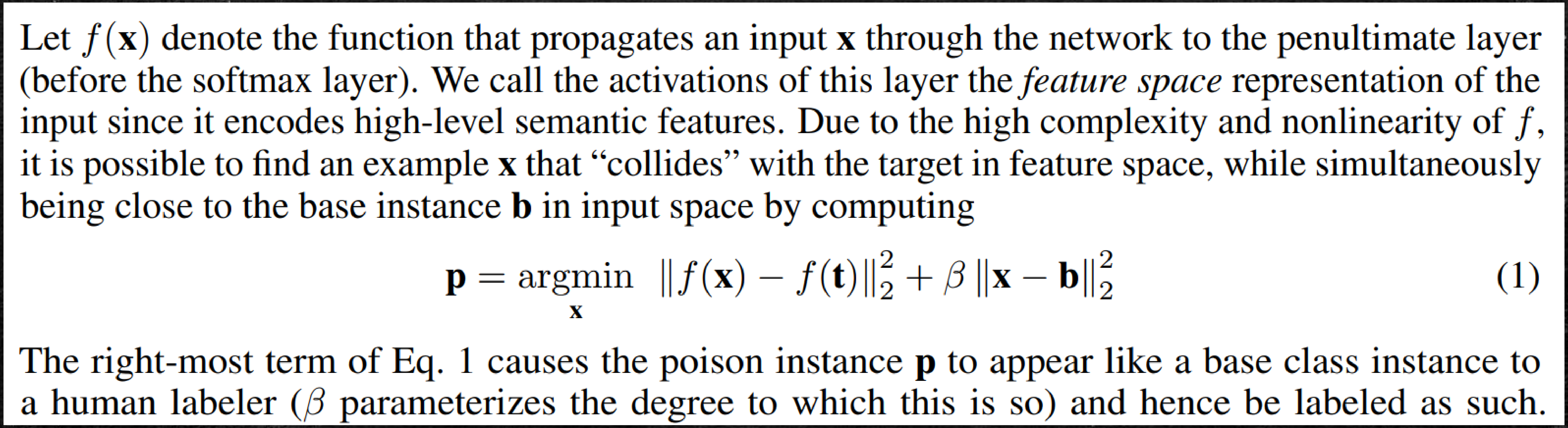


https://www.mdpi.com/2077-1312/11/6/1179

前情提要

https://github.com/ashafahi/inceptionv3-transferLearn-poison

- 我有試著看論文的 github 專案，但覺得好複雜看不懂，所以乾脆自己照著數學式寫一個



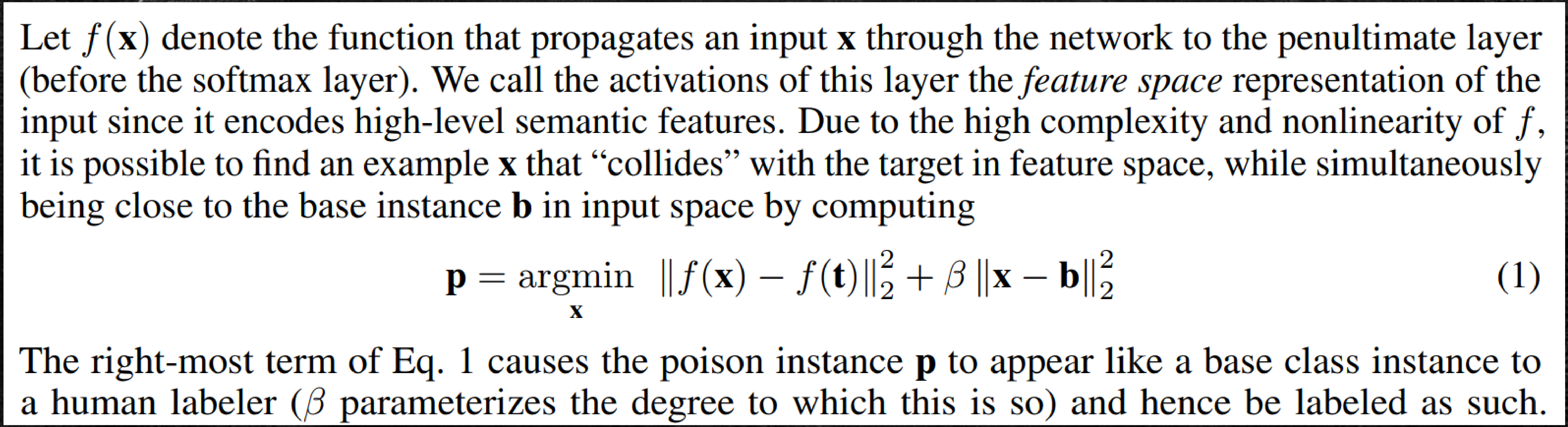
Let $f(\mathbf{x})$ denote the function that propagates an input $\mathbf{x}$ through the network to the penultimate layer (before the softmax layer). We call the activations of this layer the *feature space* representation of the input since it encodes high-level semantic features. Due to the high complexity and nonlinearity of f, it is possible to find an example $\mathbf{x}$ that "collides" with the target in feature space, while simultaneously being close to the base instance $\mathbf{b}$ in input space by computing

$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2 \tag{1}$$

The right-most term of Eq. 1 causes the poison instance $\mathbf{p}$ to appear like a base class instance to a human labeler (β parameterizes the degree to which this is so) and hence be labeled as such.

程式實作

- 可以用當初計算 CW 的方式求函數的極值
- 但是那個 f(x) 要怎麼解決?



Let $f(\mathbf{x})$ denote the function that propagates an input $\mathbf{x}$ through the network to the penultimate layer (before the softmax layer). We call the activations of this layer the *feature space* representation of the input since it encodes high-level semantic features. Due to the high complexity and nonlinearity of f, it is possible to find an example $\mathbf{x}$ that "collides" with the target in feature space, while simultaneously being close to the base instance $\mathbf{b}$ in input space by computing

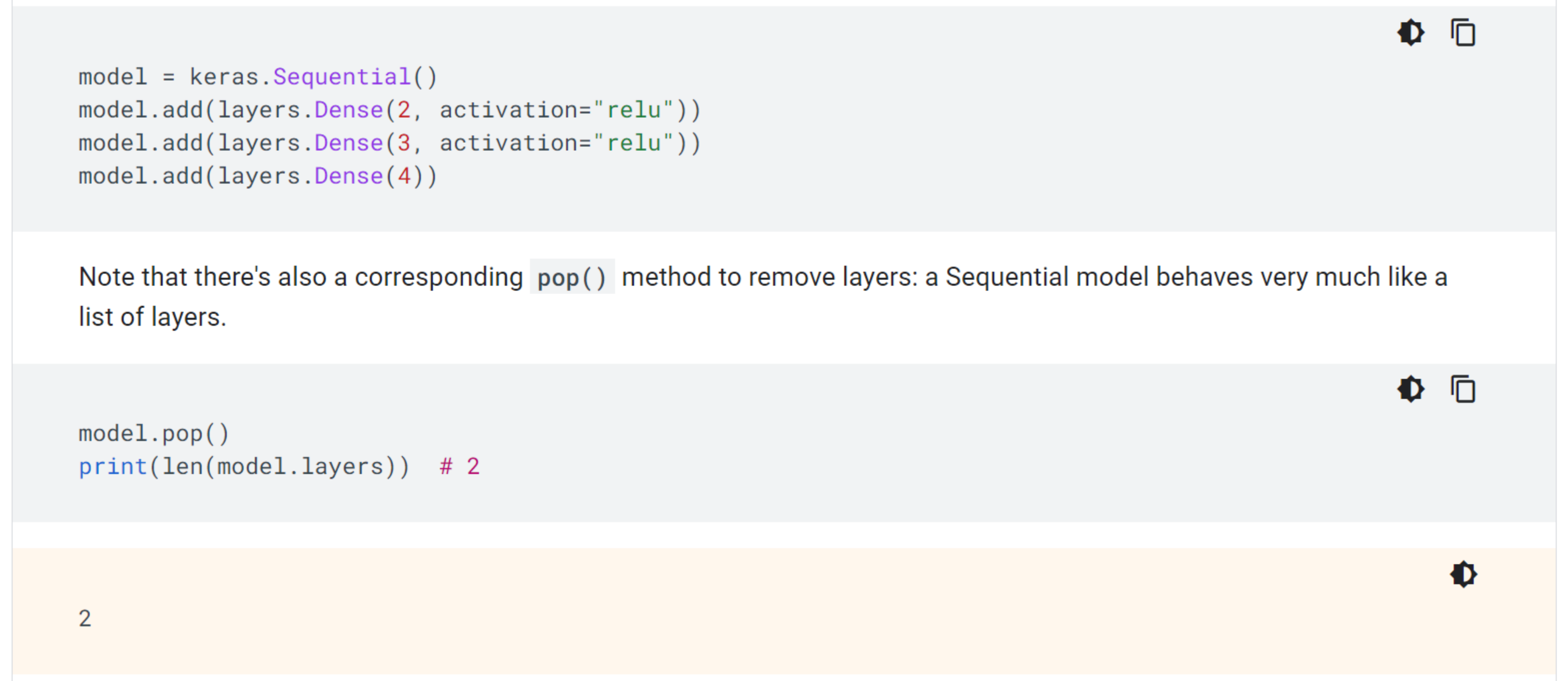
$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \ \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2 \tag{1}$$

The right-most term of Eq. 1 causes the poison instance $\mathbf{p}$ to appear like a base class instance to a human labeler (β parameterizes the degree to which this is so) and hence be labeled as such.

程式實作

https://www.tensorflow.org/guide/keras/sequential_model

- The Sequential model



```
model = keras.Sequential()
model.add(layers.Dense(2, activation="relu"))
model.add(layers.Dense(3, activation="relu"))
model.add(layers.Dense(4))
```

Note that there's also a corresponding `pop()` method to remove layers: a Sequential model behaves very much like a list of layers.

```
model.pop()
print(len(model.layers))  # 2
```

```
2
```

結論

- Clean Label 的實作雖然只是個最佳化問題，但在做的過程中也是嚇出我一身冷汗

- 原因是因為我沒注意到最佳化出來的數值居然有負數，導致當下輸入模型看似攻擊成功，但是其實無法儲存成真實有問題的圖片資料

- 回去重新溫習 CW 攻擊演算法才發現它的擾動數值有做範圍限制，難怪當初實作沒有這個問題