2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 27 – Clean Label Attack

大綱

- Clean Label 攻擊
  - 攻擊手法原理
  - 使用條件及時機
  - 程式實作
- 結論

Clean Label 攻擊手法原理

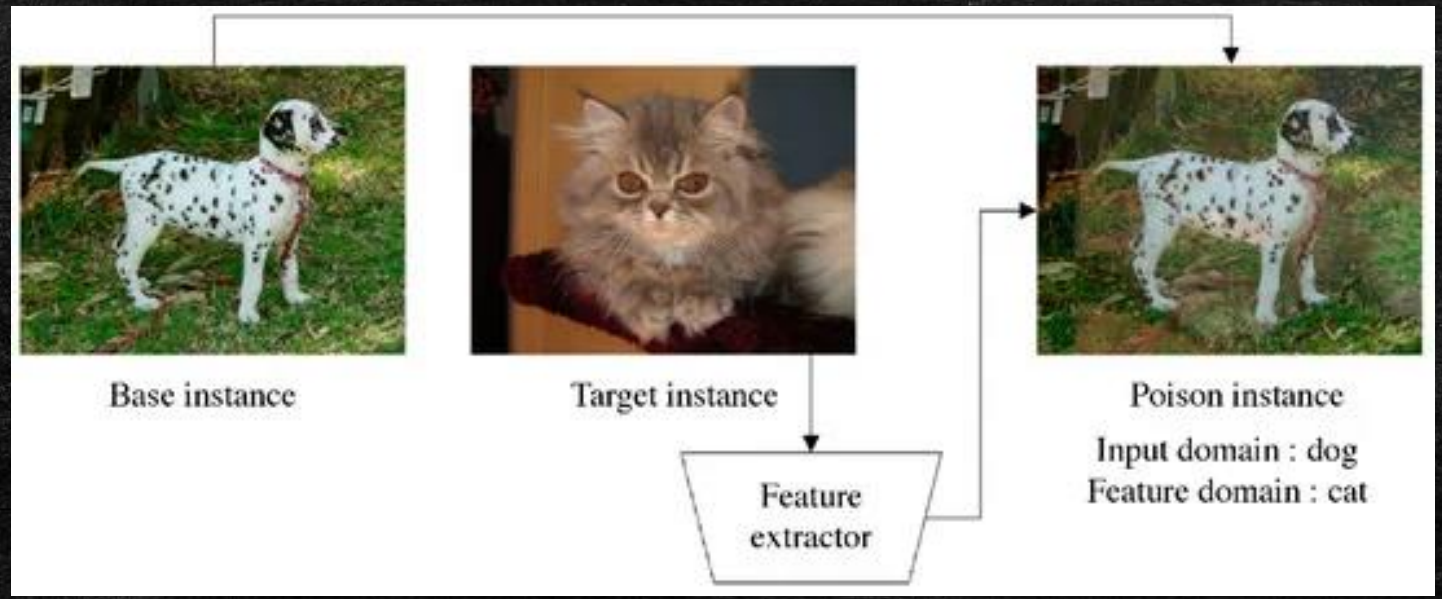
- 之前提過資料後門、模型後門，其中資料後門有些不切實際，畢竟如果是肉眼可以識別的話檢查者也不會標註錯誤的標籤上去

- 那有沒有甚麼方法可以讓檢查者看到的圖片很像是 base ，但實際上模型會判斷為是 target ?

Clean Label 攻擊手法原理

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

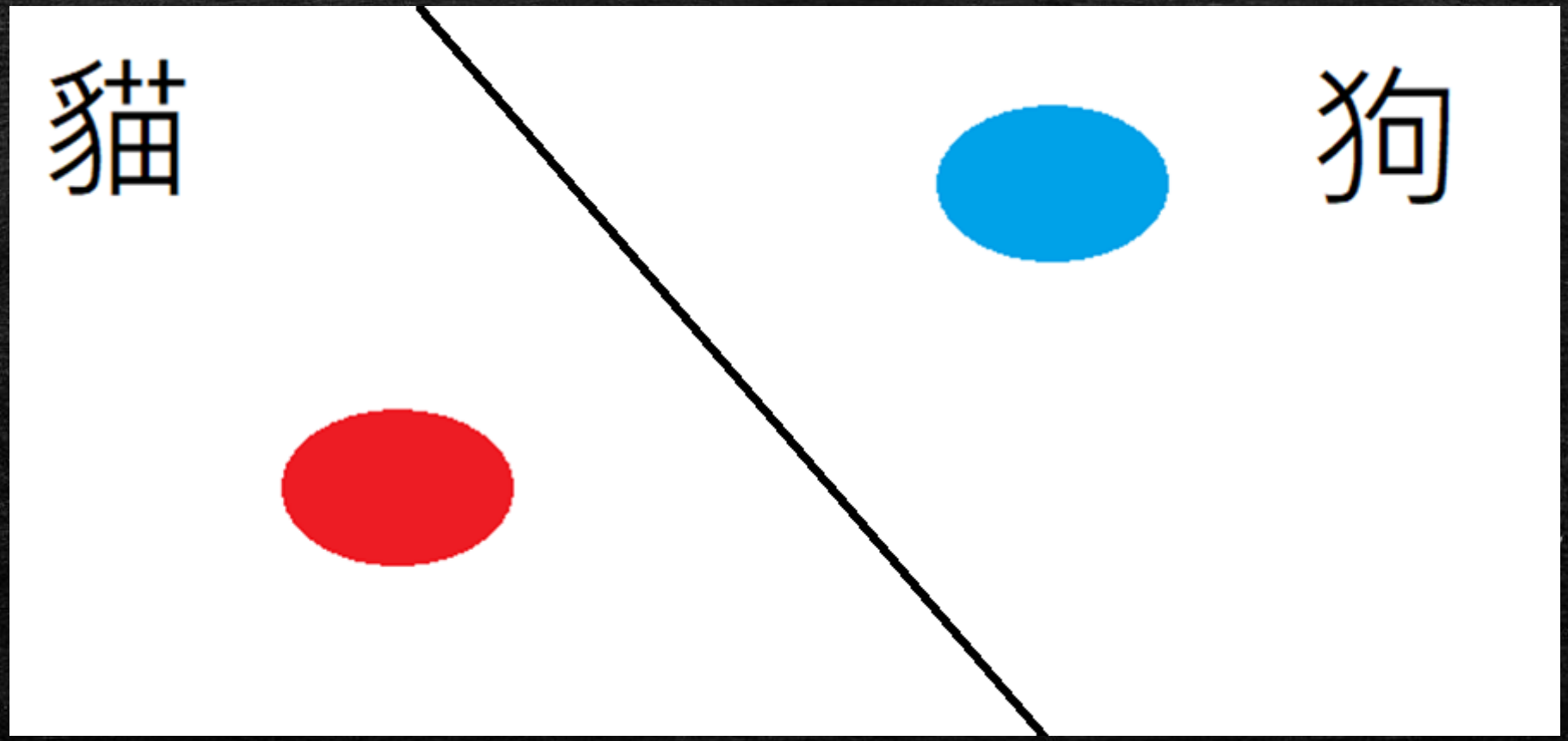


https://www.mdpi.com/2077-1312/11/6/1179

Clean Label 攻擊手法原理

https://www.youtube.com/watch?v=MLjK-SC7JSY

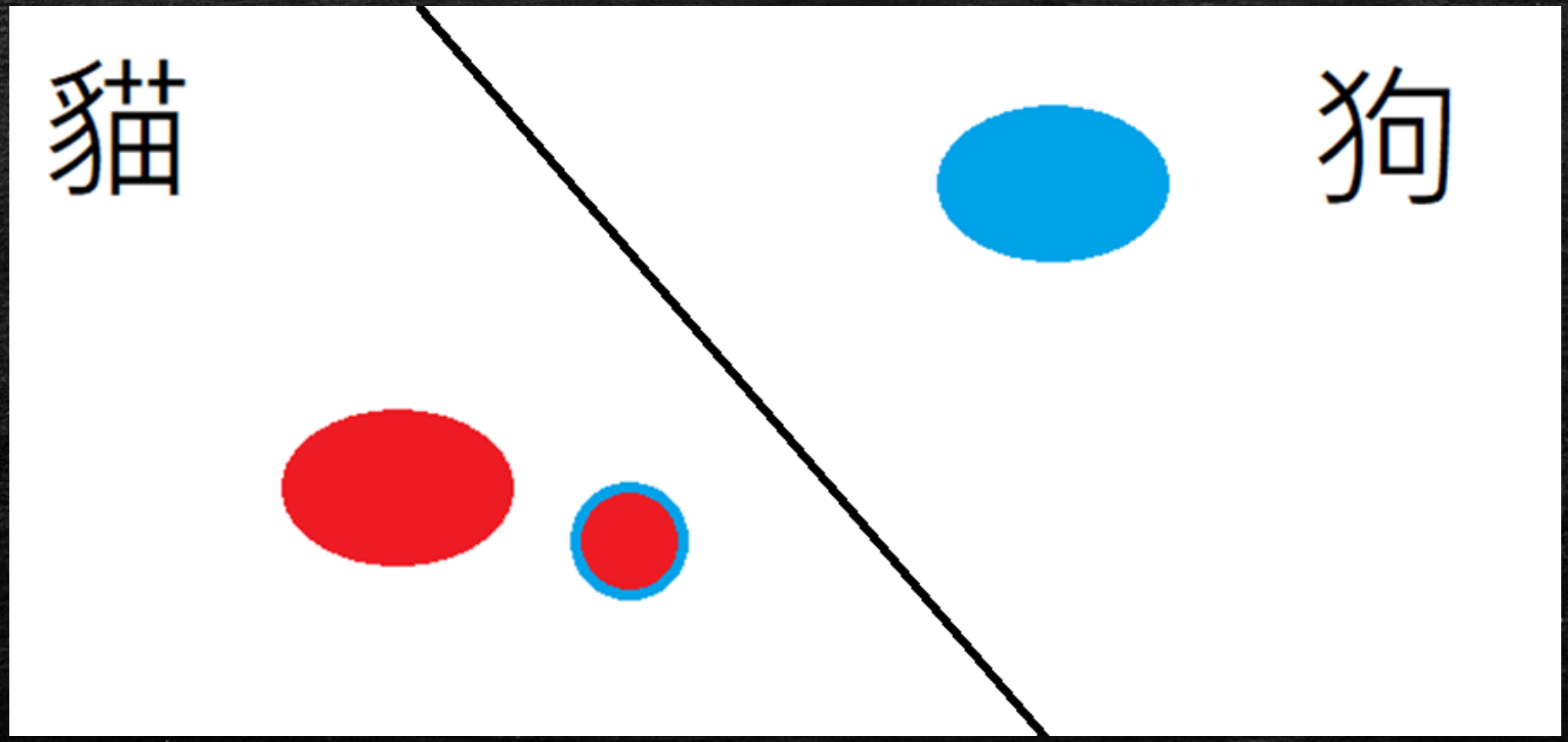
- 假設紅色是貓，藍色是狗，黑線是他們分隔線



Clean Label 攻擊手法原理

https://www.youtube.com/watch?v=MLjK-SC7JSY

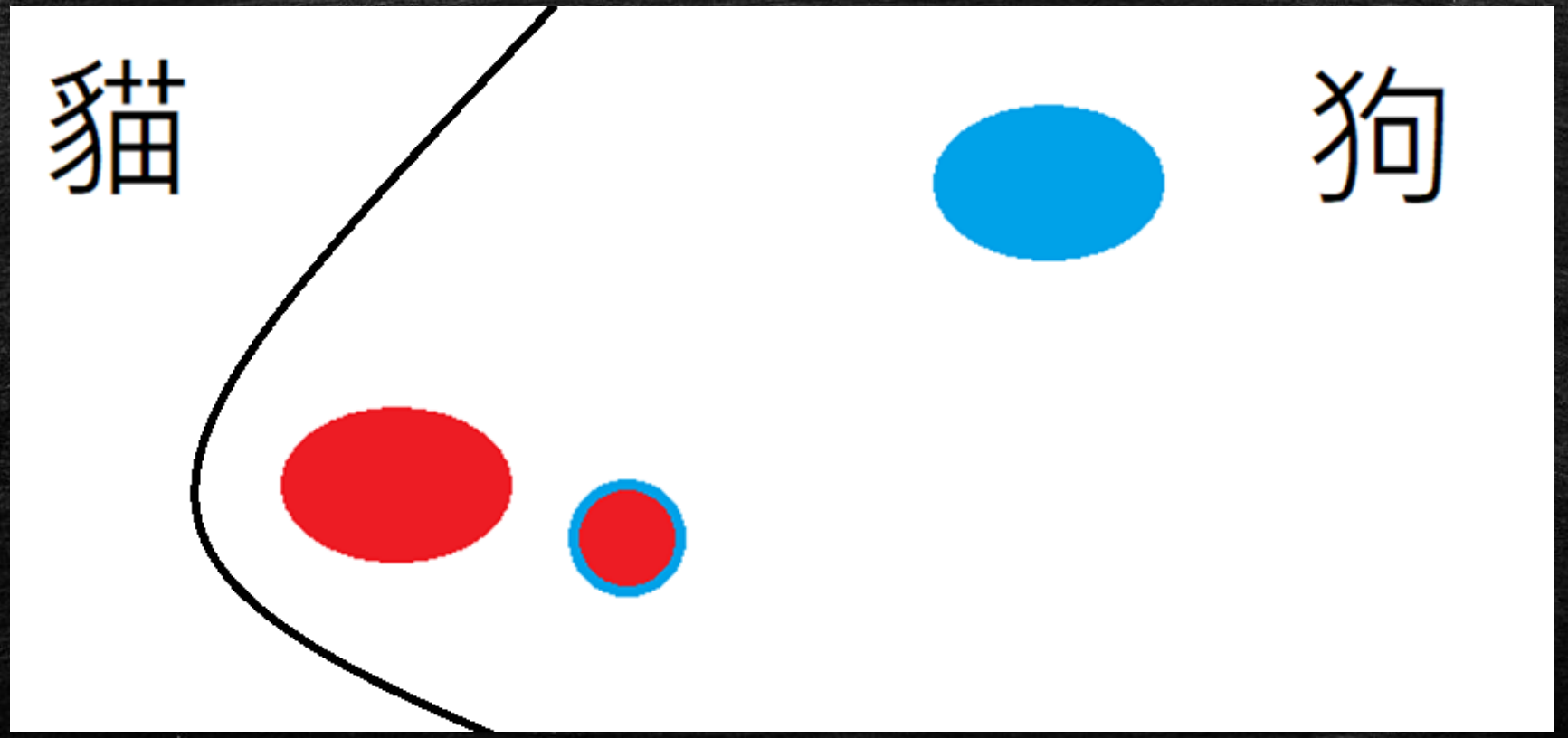
- 突然丟了一個披著狗皮的貓去做訓練



Clean Label 攻擊手法原理

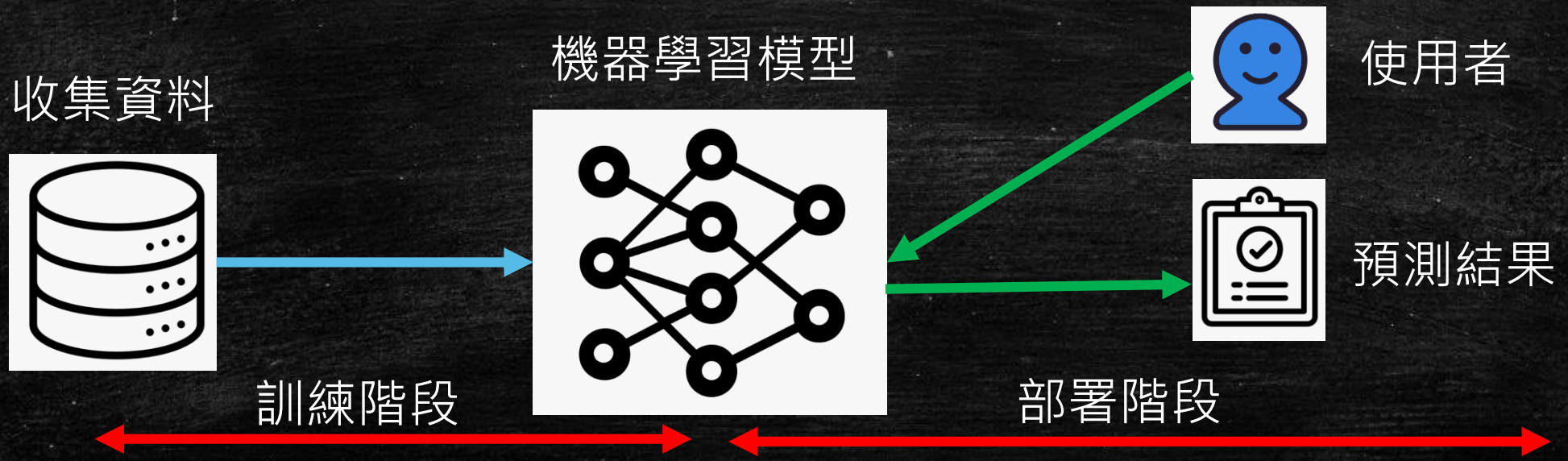
https://www.youtube.com/watch?v=MLjK-SC7JSY

- 訓練完後影響了判斷貓狗的那條線，導致原本判斷為貓的也變成狗了



使用條件及時機

- 時機點：訓練階段
- 前提：攻擊者必須能夠取得模型的訓練資料以及讀取模型
- 攻擊效果：可以讓特定資料的辨識能力下降



程式實作

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.

結論

- Clean Label 是一個相當酷的想法，利用了絕妙的設計影響了整個訓練的正確度
- 用在攻擊方面可以想像以下的情境