2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全

Day 23 – Deep Leakage from Gradients

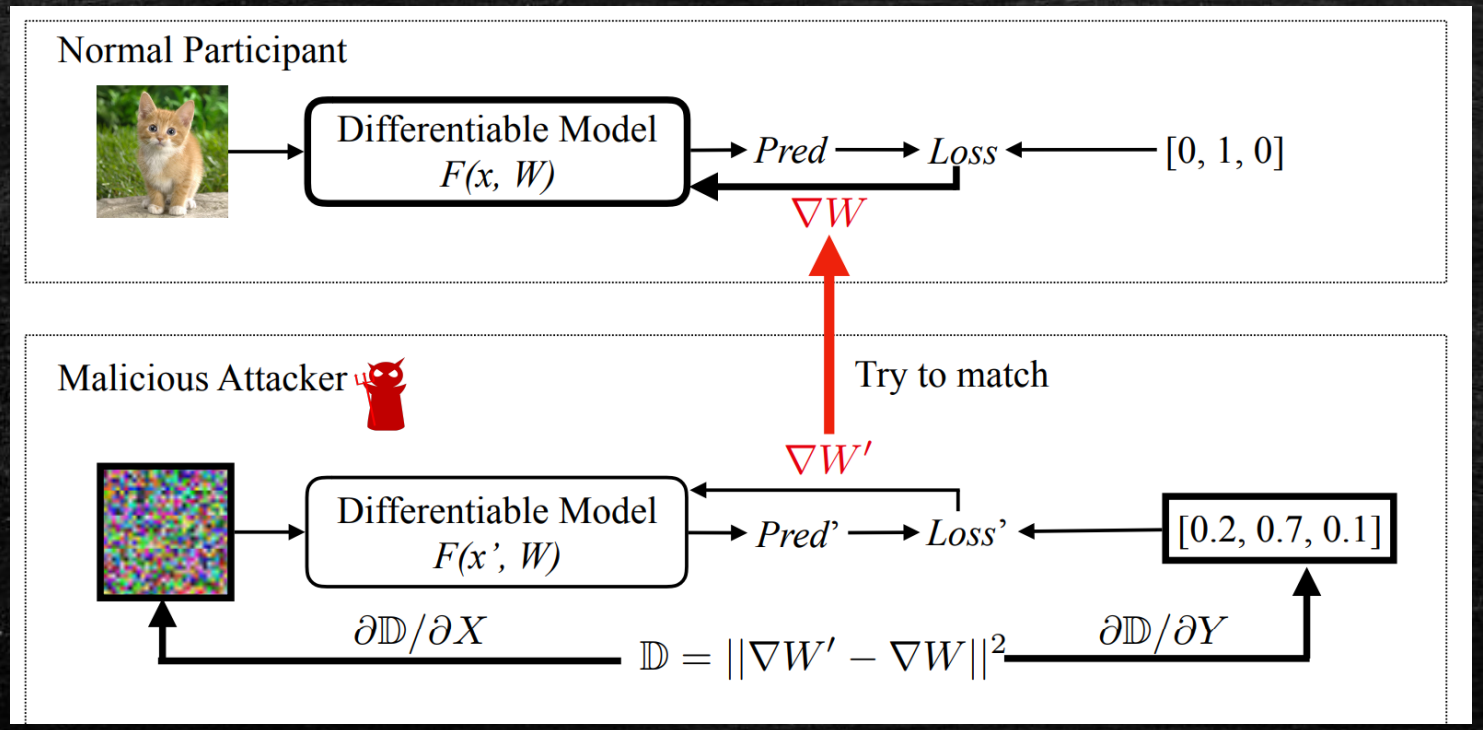
大綱

- Deep Leakage from Gradients
  - 前情提要
  - 模型設計
  - 程式實作
- 結論

前情提要

Figure 2: The overview of our DLG algorithm. Variables to be updated are marked with a bold border. While normal participants calculate ∇W to update parameter using its private training data, the malicious attacker updates its dummy inputs and labels to minimize the gradients distance. When the optimization finishes, the evil user is able to obtain the training set from honest participants.

- Deep Leakage from Gradients (2019)
  - 簡而言之，當製造出一組資料及標籤的梯度跟得到的一致時，這組資料及標籤就接近當初的輸入資料



https://arxiv.org/pdf/1906.08935

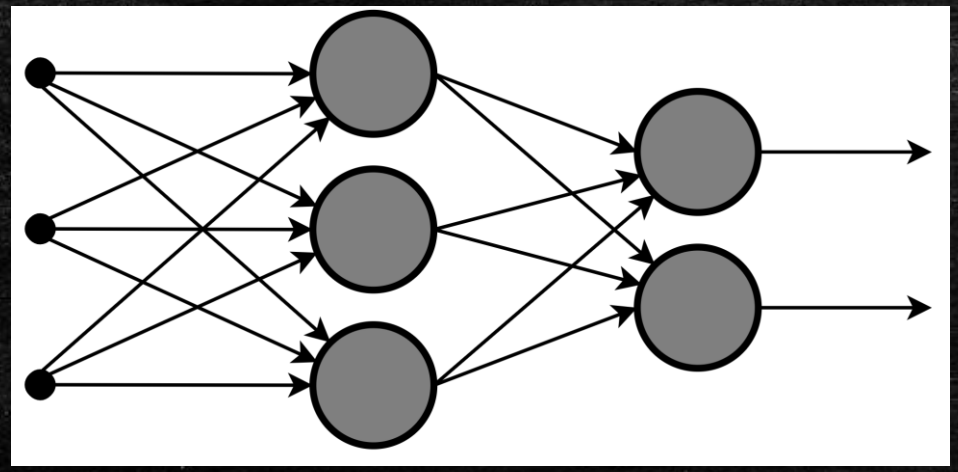
程式實作 – 參考資料 - 神之一手模型

- 這個演算法不好做，因為 loss function 變成輸入資料、標籤得到的梯度差異，然後更新資料變成回去更新輸入資料跟標籤
- 回想一下以前的模型的情況是 loss function 是預測值跟實際值的差異，然後更新資料是更新模型的權重資料
- 所以，要想辦法把輸入資料，標籤數值搞成模型權重去做更新，想法會很有趣

https://medium.com/@EastGeno/deep-leakage-from-gradients-%E5%BE%9E%E6%A2%AF%E5%BA%A6%E6%8B%BF%E5%88%B0%E4%BD%A0%E7%9A%84%E8%B3%87%E6%96%99-d23232c03bd2

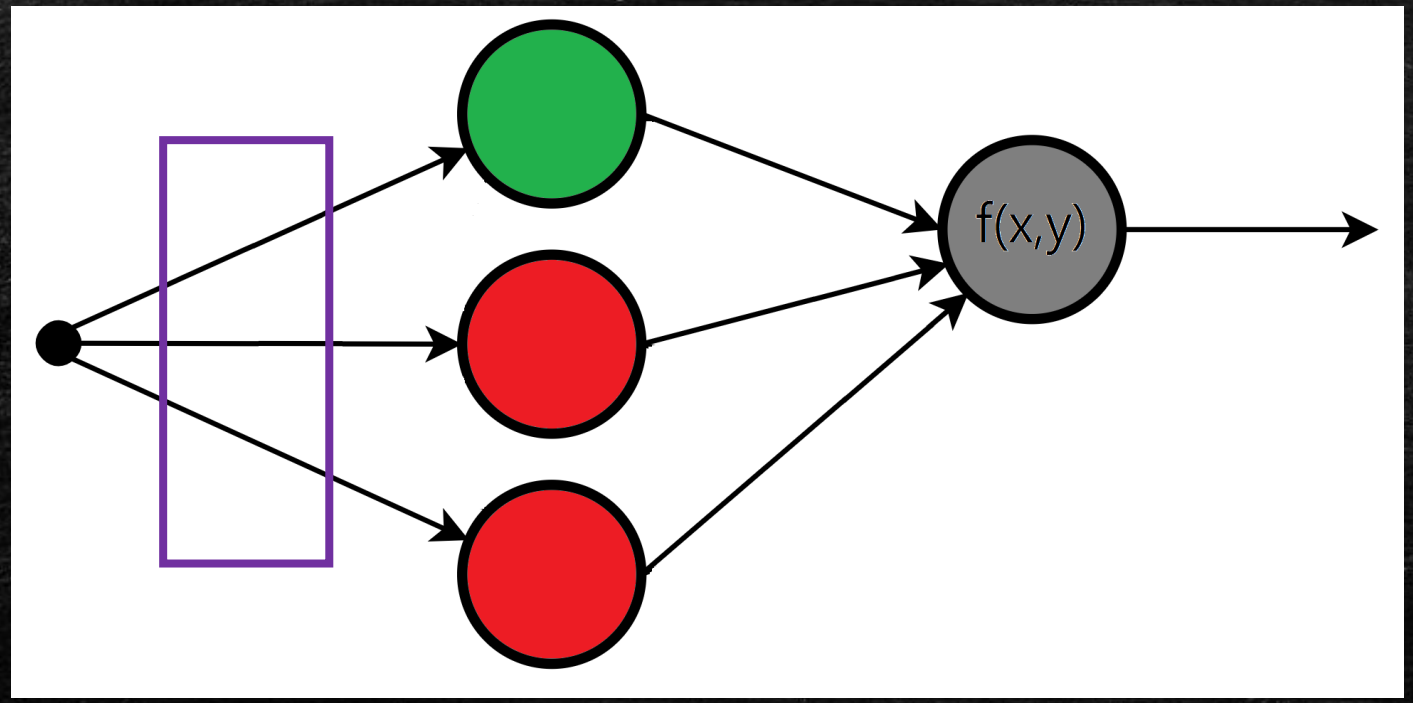
模型設計

- 上次透過 Model subclassing 的方式設計出一個模型支援回傳梯度的功能
- 但如同前頁所說，這次想要訓練的對象不再是模型參數，而是輸入圖片跟標籤
- 建立一個新的模型，把輸入圖片跟標籤的維度當作是新模型的權重，然後讓它去訓練



模型設計

- 假設綠色是輸入圖檔 x，紅色的是判斷出來的標籤 y
- 調整模型輸出值為 f(x,y) 為該資料在模型中的梯度
- 每次訓練就輸入 1 資料，讓紫色框的權重數據更新



程式實作

```
class DeepLeakage(tf.keras.Model) :
    def __init__(self,base_model) :
        super(DeepLeakage,self).__init__()
        # 宣告兩個 Dense , 一個負責圖形輸入, 一個負責判斷標籤
        self.dense1 = Dense(784, use_bias=False, kernel_initializer=tf.keras.initializers.RandomUniform(0, 1))
        self.dense2 = Dense(10, activation="softmax", use_bias=False, kernel_initializer=tf.keras.initializers.Ones())
        self.base_model = base_model

    def call(self, inputs) :
        # 這邊使用 functional API 方式讓一個輸入串接兩個 Dense
        x = self.dense1(inputs)
        y = self.dense2(inputs)
        return self.base_model.gradient(x,y)
```

結論

- 當初我看到這個模型真的覺得很神，完美的設計實作了這個 DLG 演算法
- 不過 DLG 效果並不好，後來有個 iDLG: Improved Deep Leakage from Gradients 的論文改進了它
- https://arxiv.org/abs/2001.02610