國立臺灣科技大學資訊工程系

111 學年度第 1 學期專題研究

總報告

AI 模型於對抗式攻擊的穩健性評估平台

研究組員

B10832008　蔡芸軒

B10832019　林琛琛

B10832042　鄧宥均

B10832047　楊奕儒

指導教授: 鄭欣明

中　華　民　國　112　年　1　月　6　日

目錄

1. 前言 Background & Motivation

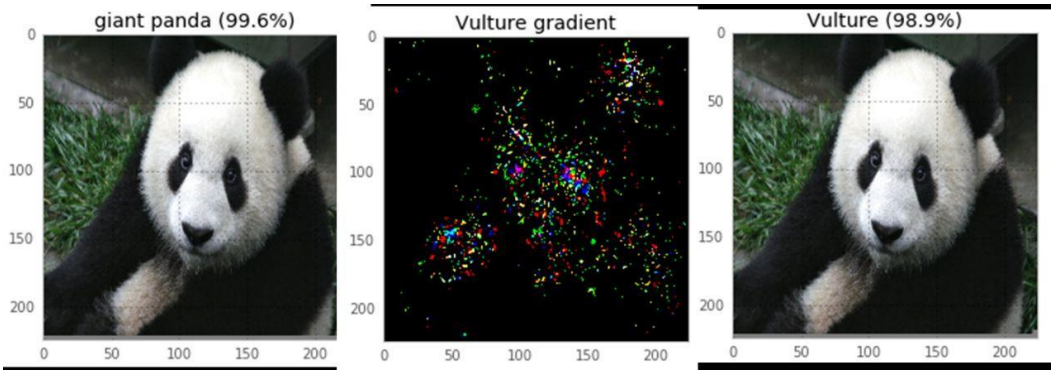
a. 什麼是對抗式攻擊 (Adversarial Attack)

對抗式攻擊是指在資安領域上的一種攻擊手法。此類攻擊往往較具有針對性，會在選定攻擊標的後，再研究目標的相關特徵，進而發起相對應的攻擊。而在 AI 領域中，攻擊者會在輸入資料上添加特定擾動，使模型做出錯誤的判斷，進而達到攻擊的目的。

b. 對抗式攻擊案例

案例一：

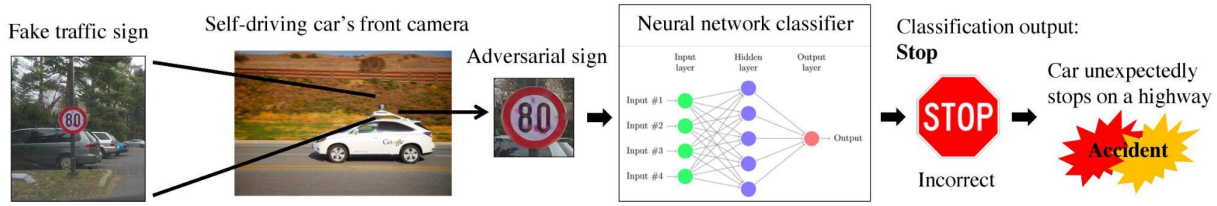
圖一為一個訓練好的圖像辨識 AI 模型，未加入擾動時，模型有 99.6% 的信心認為圖一左為大貓熊；而在將圖一左加入特定的雜訊後，可以產生出如圖一右的圖像，雖說人眼依舊可以將其辨認為熊貓，甚至無法看出其與圖一右的區別，可此時的 AI 模型卻有 98.9% 的信心認為圖一右為禿鷹。



圖一：雜訊加入前後結果對比

案例二：

近年自駕車發展迅速，許多汽車品牌都已經著手開發自駕功能，來使行車更加安全。自駕車往往透過影像辨識來判斷路標與路障，若此時自駕車所用以辨識的影像受到對抗式攻擊的擾動，很可能造成 AI 辨識錯誤，進而導致交通事故，危及駕駛者的性命。

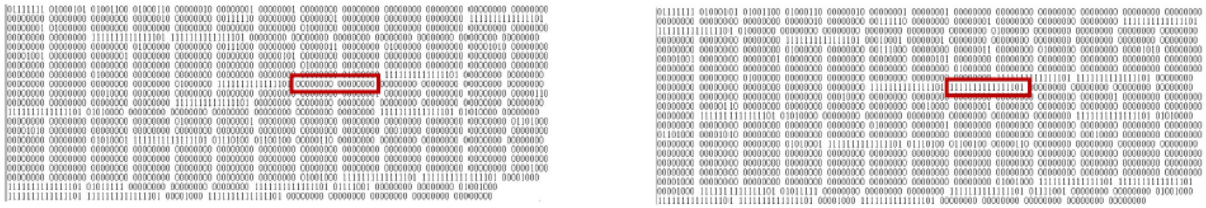


圖二：受攻擊自駕車 AI 模型示意圖

在介紹完對抗式攻擊的案例後與可能的危害後，回到本次專題的重點，也就是惡意軟體上的對抗式攻擊。惡意軟體是指以資訊竊取、系統破壞，奪取電腦控制權為目的之程式。相比其他領域的攻擊，惡意軟體的攻擊在實際生活場景中，往往會帶來更為嚴重的財產損失。舉例來說，如果個人電腦受到惡意軟體的侵害，惡意軟體開發者便有機會加密使用者的重要文件，進而勒索使用者，使使用者必須支付一定金額來解鎖其重要文件；又或者以銀行業為目標的惡意軟體開發者，若能利用惡意軟體入侵銀行，並隨意轉出及轉入資金，便會使該銀行及所有的銀行使用者蒙受鉅額的財產損失，甚至導致銀行的信任危機。

隨著惡意軟體的蓬勃發展，每個月都會有數以百萬計的惡意軟體被製造出來，使用 AI 模型對大量的惡意軟體進行辨識也因此成為必然。此時，惡意軟體開發者便會想方設法繞過 AI 模型，進而達到其攻擊目的，其中一種繞過 AI 模型的方式便是利用對抗式攻擊。在惡意軟體上的對抗式攻擊，通常是指對抗式樣本 (Adversarial Examples)。對抗式樣本會透過不影響原始惡意軟體惡意性的方式，對惡意軟體加入擾動，使 AI 模型將擾動過後的惡意軟體判定為非惡意軟體，進而實現後續攻擊的一種對抗式攻擊。

生活中，資訊安全的保護大多依賴系統和瀏覽器的防毒軟體。這些防毒軟體通常會內建 AI 模型，用以判斷一個程式是否帶有惡意。惡意軟體開發者常會在對這些防毒軟體進行研究過後，生成相對應的對抗式樣本，來對使用者的電腦進行侵害。對抗式樣本的擾動通常十分隱蔽，容易受到忽略。圖三為擾動加入前後之執行檔對比圖，可以看到擾動前後，執行檔的內容確實有所不同，但取其逆向過後的結果，程式碼卻沒有差異。



圖三：擾動加入前後對比圖

c. 動機 Motivation

由於 AI 的技術日新月異，AI 也被廣泛應用於人們的日常生活中。在資訊安全領域，AI 也逐漸被使用在商用的防毒軟體上，稱為惡意軟體檢測器，又或者 AI 檢測器。倘若 AI 檢測器無法正確分辨出惡意軟體，極有可能導致個人、企業甚至是國家的財產損失，也因此相關的資安議題應該受到更多重視。

AI 檢測器的穩健性往往需要經由對抗式樣本來驗證，又或者使用對抗式樣本來進行對抗式訓練以提升其穩健性；然而，對抗式樣本的製作在現實場景中經常受到選擇性的忽略。由於對抗式樣本具有強烈的模型針對性，且須有強大的背景知識支撐才能成功製作，在人力與物力的成本上都有高度的要求，在快速變化的商業市場上，模型開發者往往沒有餘力或者沒有能力製作對抗式樣本，導致對抗式攻擊成為 AI 檢測器的隱憂。

有鑑於此，我們嘗試開發一個惡意軟體檢測器的穩健性評估平台，藉此提供開發者快速檢驗模型穩健性的管道。希望能提升 AI 的安全性和人類對 AI 相關產品的信任度。

2. 貢獻 Contribution

在本專題中，我們開發一個平台，提供對抗式攻擊樣本(Adversarial Example) 和安全的檢測環境，讓開發者只需要幾個簡單步驟，就可以快速地評估其模型的穩健性，檢視模型的潛在問題，並加以改善，使開發者可以針對未來可能發生的狀況，進行預期性地排除。

我們的平台以網站的形式提供一般資料集和對抗式樣本資料集，讓使用者可以上傳 AI 檢測器。使用者將 AI 檢測器上傳至平台後，平台會自動建制環境，並使用我們提供的資料集進行預測，預測結果會在預測流程結束後展示給使用者。通過該結果，能讓使用者得知其 AI 檢測器對於我們所提供的對抗式樣本的防禦能力，從而再次審視該 AI 檢測器是否有進一步的提升空間。

我們期望此平台不僅能為 AI 模型穩健性評估提供方便，更能引起模型開發者對於 AI 資安議題的重視。在提高使用者對對抗式攻擊的防禦意識之餘，也能透過蒐集使用者們上傳的檢測器，進一步研究防禦對抗式攻擊的方法。

3. 實作方法 Method

a. 對抗式樣本生成

AI 檢測器通常分為動態分析與靜態分析等兩種特徵提取方式。動態分析是透過建置隔離環境來實際運行程式，並分析程式執行過程中的各項行為，來判斷其是否為惡意軟體。靜態分析則是在不執行程式的情況下，透過逆向工程提取特徵作為依據，從而判斷該執行檔是否為惡意軟體。具體來說，靜態分析可以透過反組譯反編譯等逆向手法，將可執行檔轉換為組合語言甚至更為高階的語言，並利用其含有的資訊，進行模型訓練，從而產生 AI 檢測器。

對抗式樣本的製作需同時滿足以下兩個必要條件：

1. 讓修改過的惡意軟體盡可能被模型誤判成非惡意軟體，或降低模型判斷結果的信心 (confidence)。
2. 程式被修改後仍保有其惡意性和功能完整性。

本次專題使用的對抗式樣本，是針對靜態分析的攻擊。為了滿足上述兩項條件，我們選定了不影響程式執行的位置，進行無意義內容的插入，使該對抗式樣本能夠成功繞過 AI 檢測器。主要原理為改動 main() 的起始執行位置，進而使程式在真正執行 main() 之前，進行多次的無意義跳轉 (jump)，最終才跳轉至 main() 執行惡意軟體的真正內容。

b. 環境部署

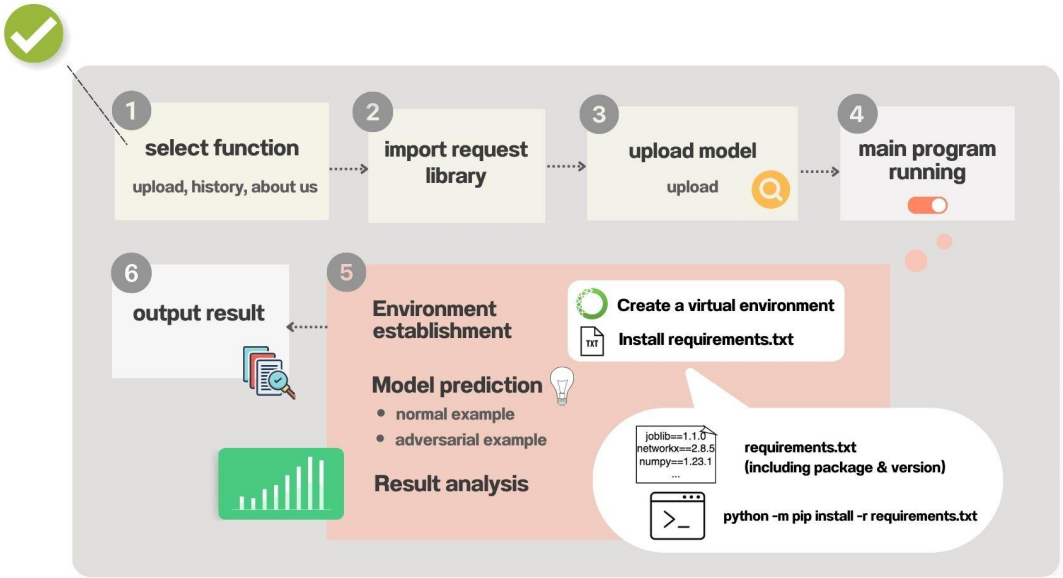
平台後端使用 Python 和 Flask 框架，在接收到使用者的上傳請求後，會為使用者上傳的 AI 檢測器建置一個新的虛擬環境，並根據使用者所提供的 requirements 檔案設置環境，並安裝所需要的套件。該虛擬環境會在 AI 檢測器預測結束後自動消滅。

c. 結果回傳

在 AI 檢測器將我們所提供的資料集全部預測完畢後，其預測成果便會記錄於平台後端的資料庫中，並以表格的形式呈現在平台前端，以供使用者進行後續的利用。

4. 結果 Result

a. 系統運作流程



圖四：系統運作流程圖

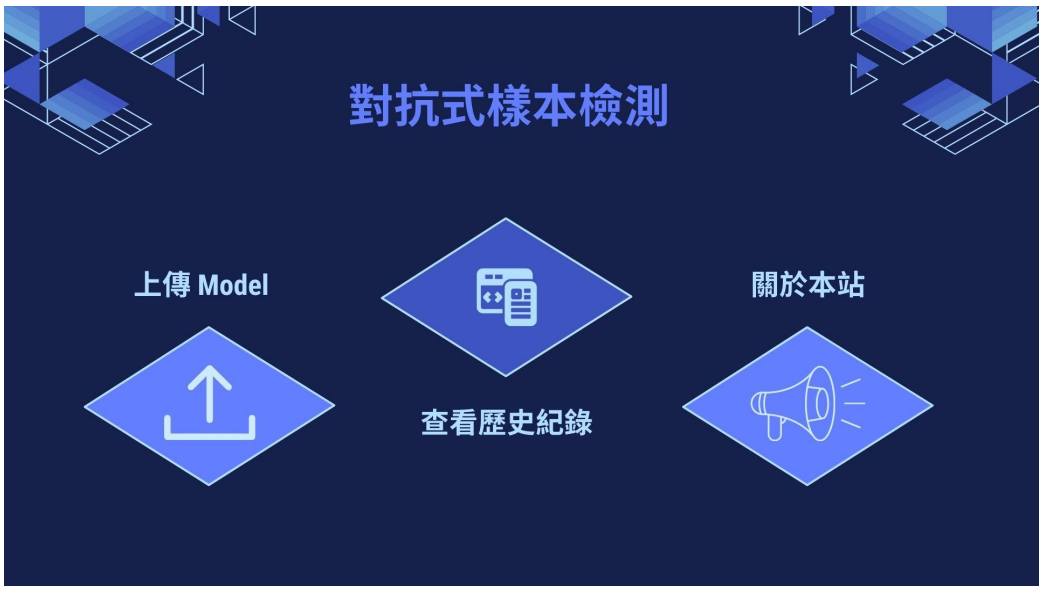
圖四為平台完整流程圖。使用者需先閱讀上傳格式規範，並依照規範調整欲上傳的 AI 檢測器細節，再將檢測器上傳至平台。同時平台會根據各個使用者的使用工具與環境需求，建置相對應的虛擬環境。

在虛擬環境建置好後，便會於其中執行該 AI 檢測器，以預測本平台提供的資料集，並將預測結果展示於如圖五的示意網頁前端頁面。預測結果包含資料集之混淆矩陣、Accuracy、Recall、Precision、F1 Score 等評估指標。

這些數值能讓使用者參照不同指標，進而調整其檢測器，以降低檢測器受對抗式攻擊的風險，並提高其穩健性。

圖五：評估指標示意圖

b. 實際運行範例



圖六：系統首頁

圖六為系統首頁，點擊首頁最左邊「上傳 Model」圖示後，便會進到上傳規則頁面，如圖七。使用者須閱讀規則並點擊「我同意上述規則」， 接著點擊 next 按鈕，方可進入上傳頁面。

圖七：上傳規則

圖八為上傳頁面，使用者須點擊「選擇檔案」方框，選擇欲上傳的檢測器後，點擊 submit 按鈕將檢測器上傳到平台以執行後續的評估流程。

圖八：上傳頁面

檢測器上傳到平台後，平台會以上傳的檢測器預測平台所提供的所有可執行檔，並將預測結果展於在圖九的結果頁面。結果頁面有四種評估指標供使用者參考，分別為 Accuracy、Recall、Precision、F1 Score。

圖九：結果頁面

5. 結語

現代社會中，利用 AI 模型來解決大量且重複的分類工作，能大大地便利我們的生活，但 AI 模型的可信度卻也容易因對抗式攻擊而受到質疑。本專題以惡意軟體 AI 檢測器的對抗式攻擊為主軸，提供了一個可檢驗 AI 檢測器穩健性的平台，能使模型開發者快速檢驗自身檢測器的不足之處，並能夠藉此進行修正。

然而對抗式攻擊的對象並不僅侷限於此，諸如音訊或影像辨識的 AI 模型也是潛在的攻擊對象。而我們選擇了惡意軟體這一相對冷門的領域，是因為相比起其他領域，錯判惡意軟體，會在現實面上帶來更大的危害。我們希望能以此平台作為契機，讓模型開發者們意識到對抗式攻擊的潛在危害。

此外，本平台期許能在未來進行擴充，在模型的規範上提供更多的彈性，在平台的功能上提供更多的選擇，以及在資料集上提供更多元的內容，使更多的 AI 檢測器能夠藉由本平台來評估其穩健性，使本平台能夠成為未來 AI 科技中，一道堅實的護盾。