經濟部 109 年度
物聯網系統檢測與驗證計畫
合作研究計畫

《機器學習模型資安滲透測試技術》

建議書徵求文件

財團法人資訊工業策進會

中華民國　109 年 03 月

109 年度合作研究計畫建議書徵求文件

一、簡介

財團法人資訊工業策進會(以下簡稱本會)所屬資安科技研究所(以下簡稱本所)執行經濟部「物聯網系統檢測與驗證計畫」計畫，在物聯網與AI整合應用近年快速發展趨勢下，AI技術如今已被廣泛整合、應用於多方領域之中，其中包括基於AI的惡意程式偵測系統，本計畫將致力於研究Adversarial AI on malware detection相關題目。

對抗攻擊(Adversarial Attack)原本是用來攻擊AI模型的演算法，能夠在未取得神經網路相關參數的前提下，添加少量雜訊於深度神經網路上使得AI model產生誤判，該技術多用於圖片、影片上，但隨本計劃研讀國外最新研究發現，有許多論文將該技術領域由圖片投射至惡意程式偵測領域，透過特定演算法以產生能夠使現今的惡意程式偵測系統發生誤判成正常程式而鑽過漏洞，因此，本計劃欲建構、實作生產adversarial example for AI malware detector與生產Adversarial Pattern的自動化系統，希望能夠藉此評估AI malware detector的安全性。

二、計畫目標

隨著AI(Artificial Intelligence)近年來不斷發展，越來越多領域與AI技術相結合。譬如AI影像辨識、AI醫療輔助系統、AI工業系統等等，近年來，AI也開始被用於惡意程式偵測上，然而，AI技術正面臨對抗例樣本(Adversarial Example)的重大考驗，尤其當AI用於資訊安全領域時，對抗例樣本欺騙深度神經網路的能力將成為一大威脅。

具體而言，本計畫不僅針對AI的攻擊演算法進行研究，還可以將其實踐於真實蒐集之惡意程式資料集中以證明其可實作性與可用性。此外，本計劃還會為此設計對抗例樣本生產流程與架構自動化對抗例生產系統建構對抗例樣本，以評估現今AI惡意程式偵測技術背後AI影像辨識模型的安全性。

研究並實踐對抗例樣本後將會建構一公開測試網站平台，對於資安技術開發有興趣的人員，可以透過本平台比較不同模型和對抗例攻擊之間的防禦程度，並使用本平台作為日後開發資安防護相關技術的依據之一。

本計劃目標在於針對AI惡意程式偵測系統，建構binary domain的惡意程式對抗例樣本與自動化對抗例生產系統，一方面能讓國內AI服務提供業者評估自己的AI惡意偵測系統之實用性與安全性，以期建立更穩固的AI應用。

本計畫目標亦包含讓AI資安防護的開發者可以透過本計畫所開發的網站，瞭解現有AI模型的可依賴度，並能夠自行選擇欲比較的對抗例攻擊方法和AI模型來評估，並決定防禦方案以及早面對模型可能面臨的潛在威脅，預防惡意人士惡植的誤導攻擊例。此外本計畫所建立的網站所包含的研究數據等，也可以作為AI資安防護的技術開發參考，以協助未來能開發出更堅強的AI資安防護技術。

三、計畫範圍

為保護AI使用者的個人資訊安全與提升其資安意識，本計畫擬研究開發real time AI對抗例樣本的生產流程與實作自動化攻擊系統，研究範圍如下：

1. AI對抗攻擊例生成方法之研究：

本計畫研究者須先了解針對binary的對抗例攻擊演算法及其應用範圍，以期有效率地以ELF格式產生對抗例樣本。

2. AI惡意程式偵測之研究：

本計畫研究AI惡意程式偵測系統相關技術並投入真實蒐集之惡意程式資料以測試其演算法、系統架構之可行性與實用性。

3. 設計對抗樣本生產流程與自動化生產系統：

本計畫主要目標，須結合上述技術設計自動化對抗例樣本生產流程與系統架構，以此實作具攻擊AI惡意程式偵測系統的binary對抗例樣本。

4. 設計AI模型可靠度分享平台：

本計畫比較了多種基於圖片的對抗例攻擊種類以及AI模型，將運用這些測試數據建構出完整的比較系統並且建立在網站上，以期能將測試比較之結果讓資安相關，開發人員透過此網站來測試模型對於不同攻擊例之可靠度，作為日後開發之依據並促進資安開發技術之進步。

四、預期成果(明確說明合作研究成果之產出)

1. AI模型攻擊與自動化攻擊系統研究及可靠度分享平台技術之期中及期末報告
3. 自動化對抗樣本生產系統之程式原始碼，包含：
   a) AI惡意程式偵測器
      - AI惡意程式偵測器之使用說明文件
      - AI惡意程式偵測器之程式原始碼
   b) AI對抗樣本生產模組
      - ELF binary對抗例樣本生產流程圖及模組使用說明文件
      - 合成ELF binary對抗例樣本之原始碼
4. AI模型可靠度分享平台之程式原始碼
      - 基於 3 種資料集(包含MNIST、CIFAR10 及ImageNet等)，整合 6 種AI模型(包含VGG19、ResNet50 及InceptionV3 等)，與至少 10 種基於圖片的對抗例攻擊型態之流程圖及說明文件
      - 網站平台設計模組之原始碼
      - 分享平台中對抗例圖片預覽和比較數據圖模組之原始碼

※前述成果如有專利構想或專利申請產出時，需注意專利申請之新穎性(novelty)。因凡經公開發表之研發成果，如擬申請專利，須於公開發表後 6 個月內完成，前述成果如是以論文方式公開發表，將無法取得大陸與歐盟等國之專利。

五、執行方式(包括計畫時程、計畫分工方式)

本計畫整體規劃流程如下：

1. 分析相關文獻蒐集及研究：針對AI之對抗例生成演算法與AI惡意程式偵測技術進行研究，並蒐集AI對抗例樣本相關研究資料。
2. 系統架構設計：根據第 1 項的研究分析結果，設計攻擊對抗例之流程規範與模組化之系統架構，以及公開測試平台並匯入本計畫之研究結果以及實驗數據，讓使用者可以即時在網站上選擇欲比較的資料集和攻擊種類並測試AI模型的可靠度，產生出比較結果長條圖。
3. 系統開發環境建置：根據第 2 項相關架構設計，進行相關環境以及公開線上平台之建置。
4. 展示系統實作：根據第 2 項的架構，實作對抗例樣本之生產，並攻擊目標之AI惡意程式偵測系統，最終整合公開測試平台供使用者應用該系統模組，且能選用欲比較之對抗例攻擊得到即時回饋的可靠度資訊。

六、計畫期程及預估計畫總經費

計畫執行區間： 109 年 01 月 01 日 至 109 年 12 月 15 日
總經費：60 萬元整

七、驗收標準(含教育訓練)

109 年 12 月 15 日前完成以下項目：

1. 期中、期末驗收分別預計於 109 年 09 月 27 日與 109 年 11 月 29 日完成：
   - 期中驗收的內容包含：
     - 產生ELF格式的binary對抗攻擊例之程式原始碼
     - 生產對抗樣本系統之設計流程規範、系統架構圖
     - 系統模塊基礎建設(infrastructure)之原始程式碼
     - 網站平台之設計流程和系統架構圖
     - 網站平台使用到的對抗例圖片攻擊例之原始程式碼
     - 網站平台即時分析圖之原始程式碼
   - 期末驗收的內容包含：
     - 經調適之整合對抗例樣本生產系統
     - 系統、模組使用說明文件
     - 網站平台的完整網頁以及其程式碼

2. 進度討論會議：每月召開一次進度研討會議

八、技術能力需求(請詳述所需要之技術能力或專長)

提案團隊需符合下列資格：

1. 國內專業學術研究團隊
2. 具備實作AI Adversarial Attack演算法之經驗與能力
3. 具備使用AI malware detector技術之能力
4. 具備Machine Learning基礎知識與調適Deep Neural Network之技術
5. 具備自有整合、開發軟體工具系統之能力
6. 其他：如熟悉 C/C++、Java、Python、網頁前後端等系統程式的運作

附件1：契約書格式
1-1：計畫書格式
1-2：經費動支報表
1-3：成果報告撰寫須知
1-4：報告格式
1-5：論文格式
1-6：保密聲明書
1-7：委託匯款同意書