人工智慧影像辨識、網路入侵偵測系統安全測試方法與工具

（一） 計畫摘要

近年來隨著深度學習網路的發展，人工智慧的應用已擴及到各領域層面，包括智慧製造、智慧醫療、自駕車、網路安全及精準農業等，然而接踵而來的就是系統的穩定及安全性問題。基於機器學習的深度學習可能在訓練階段或判斷階段受到攻擊，前者故意錯置標籤(label)，讓學到的模型錯誤，稱為下毒攻擊(poison attack)，後者竄改輸入資料，造成誤判，稱為輸入攻擊(input attack)。由於深度學習的架構與應用各有不同，並沒有辦法有一個通用的模型或方法來檢驗或對抗這些攻擊。例如同樣是影象的應用，就有上百種常見的機器學習演算法，而相同的演算法可能有不同的攻擊手法。所以如何檢驗一個宣稱具備機器學習的應用是否具備防範攻擊的能力需要視應用的範疇而分別研究。本計畫預計以三年的時間選定五類應用以上的應用(如影像分類、自然語言處理-文字與語音、數據分析、網路安全)、四大類的演算法(深度神經網路、卷積神經網路、遞歸神經網路、強化學習)，分階段發展一個通用型測試框架與工具鏈。計畫成果將用於測試各種人工智慧系統的安全性與可靠性。

在第一年的計畫中，我們選定兩種應用(影像分類、網路安全)，其中影像分類再細分為一般的影像應用與在工業4.0上的應用。預計針對這三個領域的應用的攻擊，即影像分類應用攻擊、工業4.0應用攻擊與防禦及網路流量攻擊，進行研究。將對這三個應用各提出一項研究方法及測試工具，以期能在現有的深度學習攻防技術上，找尋突破性的發展，並開發相應之測試工具來應對實務上的攻防需求。

而在工業4.0應用攻擊與防禦領域，近年來，隨著科技的進步，用於工業上的技術也逐漸提升。物聯網的普及與網路速度及運算能力的提升，隨之而來的是工業4.0的時代。工業4.0的概念為將所有的機器連上網，透過網路與智慧製造賦予機器智慧，未來的製造業將會注重機器之間的連接度(Connectivity)[2]，不但能使成本降低、高品質的產出，除了提高業者的利潤之外，透過深度學習的技術將能應用在生產線上：瑕疵檢測的良率及效率，傳統的生產線大多是利用人工以及肉眼去進行瑕疵的檢測，而透過深度學習的技術檢測將能更精準的判斷產品是否有瑕疵，提高產品的品質，並且提升判別的速度。以生產鋼材為例，鋼材的生產大多用於大型建設以及交通設施上，如果因人力的判斷錯誤而導致疏失，後果將不堪設想。而透過深度學習結合影像分析，可以有效的篩選鋼材的品質，對於事後的維修成本上想必有極大的影響。而應用在機台上，也能有效的監控機器，降低機器異常帶來的損失。在工業4.0中的機台檢測結合深度學習的應用上，業者通常在機器上安裝監測裝置，並透過監測裝置收集

機台數據，最終將其透過深度學習訓練的模型偵測機台的行為是否異常，達到節省人力成本以及分類狀態等相關應用[3]。

隨著深度學習的興起，攻擊者也逐漸聚焦在深度學習的安全機制問題。由於深度學習本身的防禦機制較為薄弱，因此攻擊者逐漸將攻擊目標轉移至深度學習模型上[4]，當透過深度學習技術的機台遭到攻擊者的攻擊時，將會造成嚴重的損失，本計畫為致力於避免上述情況發生。本計畫可分成兩個部分，第一部分為針對具有深度學習能力的機台狀態識別分類系統進行攻擊，讓系統誤判其分類。第二部分為防止結合之深度學習應用遭攻擊淪陷，提出了具備防禦機制的機台狀態識別分類系統。透過第一部份的攻擊，測試系統的穩固性，讓機具及對於該系統之攻擊進行研究。而第二部分的目的是針對系統防禦性與對抗系統判別率之提升，最終能識別攻擊者的攻擊，將其無效化，以達到增進基於工業4.0應用系統之穩固性的成果。

在網路流量攻擊領域，由於近年來深度學習等技術已臻成熟，所以過去以攻擊特徵辨識為主、機器學習為輔的入侵偵測機制，也強調越來越多的機器學習能力。其中主要的原因在於各種網路攻擊變化不窮，各種像是零時差(zero-day)的攻擊、先進持續威脅(advanced persistent threat, APT)等攻擊型態，使得各種依賴既有攻擊特徵的偵測方式疲於應付，使得更加自動、具有學習能力的深度學習機制受到重視。然而如同過去攻擊者會針對各種特徵為主的入侵偵測機制進行迴避攻擊，針對機器學習的迴避攻擊方式在近年來也有諸多探討，但目前的研究大半都是與影像辨識相關。對於如何變異惡意網路流量，使得入侵偵測系統產生錯誤的判斷，但仍然保持一定的攻擊效能，這方面的研究仍然相當有限，也缺乏實際的網路流量變異工具進行實地驗證。這也是我們在這個計畫中需要去著手進行的地方。

（二） 研究計畫之背景

現今工業 4.0 與智慧製造為製造產業鏈的主要發展趨勢，其中的工業物聯網與大數據分析為主要重點。工業物聯網技術為收集工廠中各樣的機台的監測數據以及生產線環境資訊，並將各項數據透過網路傳輸，藉由深度學習進行大數據分析，透過訓練出的模型可使其更貼近生產環境狀況及機具的狀態，並藉此提高生產效能。

在工業 4.0 結合深度學習技術上，如文獻[5][6][7][8][9]皆使用卷積神經網路(Convolutional Neural Network，CNN)應用於預測機台是否故障或監控器具磨損狀態及預測器具使用壽命，這些應用大大的提升機器的使用效率及產能。上述在機具狀態資料集上分別透過傳感器波形、電流或信號轉換成圖像後，再經由 CNN 對圖像進行特徵提取及分類，最終皆取得不錯的成效。其中文獻[5]所使用的模型更是以 CNN 的變形版本 AlexNet 進行研究。

在深度學習中，由於模型容易受到惡意者攻擊，其攻擊可分為兩種模式: model poison attack 與 data poison attack，前者的攻擊較容易被偵測，因模型中的變動參數是顯而易見的，而後者是根據 model 的 input 需求刻意設計錯置 label，因此較難被發現。然而，以現階段防禦技術來說，大多是針對 model poison attack 進行防禦，而 data poison attack 之防禦由於偵測不易，因此尚未成熟。 因此，在工業 4.0 產業日漸興盛下，結合上述工業 4.0 與運用深度學習提升產能的趨勢，在深度學習中如何進行有效的攻擊而不被偵測出來，以及如何針對攻擊做出有效防禦，為一重要的議題。

（三） 國內外產業需求背景

在工業 4.0 發展趨勢下，產業界紛紛投入研究。德國為首先提出工業 4.0 概念者，在各方面技術已發展成熟，並主打智慧化與虛擬化，取得國內外許多的訂單，為工業 4.0 的龍頭。美國將工業 4.0 稱作「工業互聯網」，並注重在機具間的連結性與數據間的互通性，著力在大數據以及雲端計算部份。中國將其稱作「中國製造 2025」，雖然中國在技術方面較不成熟，但藉著人力資源多，以及擁有全球最大的市場，並搭配政府的政策實行，想必會有突破性的進展。在台灣方面，面臨到的的問題為少子化與高齡化，以及品質不穩定的問題。許多產業紛紛採用機器人以及機器手臂等自動化的技術，在工業物聯網的大潮流下，將設備連上網後，深度學習的技術也日漸重要，納入深度學習方法提升產能為一重點目標，但基於深度學習安全性上的考量，若使用不具備防禦機制的系統，容易成為攻擊者攻擊的目標，造成整間工廠癱瘓，其損失將不堪設想。因此工廠想加入深度學習的技術，須具備能夠防禦的技術，避免被攻擊成功導致淪陷的可能性。因此，本計畫預計產出一項機具的狀態識別系統，並且使用生成對抗網路技術對該系統進行攻擊，藉此調整與增進該系統防禦功效，改善工業 4.0 之深度學習安全性。