結合語言模型與特徵機制之整合式網路入侵偵測告警系統

專題編號：113-CSIE-S032
執行期限：112 年第 1 學期至 113 年第 1 學期
指導教授：孫勤昱
專題參與人員： 110590006 顏睿寬
110590024 許宸瑋
110590029 陳思群

一、摘要

隨著網路科技的快速發展，網路安全威脅也日益增加，例如 DDoS 和 SQL 注入攻擊 [1] 等。為因應這些挑戰，本研究利用 CIC-IDS2017 資料集 [2] 訓練一個自然語言模型，以分辨網路流量中的惡意攻擊。我們將結合 Wireshark 定期監控網路封包，經過處理後將資料傳進訓練好的模型進行判讀。一旦檢測到惡意封包，系統將立即向管理員發送告警訊息，並提供惡意流量的來源及相關資訊，以利及時採取必要的防護措施。我們期望透過一系列自動化流程能為網路安全領域帶來更有力的防護支援。

關鍵詞：網路安全、入侵檢測系統、深度學習、語言模型、CIC-IDS2017

二、緣由與目的

隨著網路攻擊手法日益複雜和隱匿，傳統的網路安全防護措施如防火牆、入侵檢測系統 [3] 等已經難以有效應對新型態的網路威脅。尤其是零時差漏洞攻擊這類利用系統或應用程序未知漏洞進行的攻擊，對傳統防護手段更是一大挑戰。

因此，我們期望能透過機器學習和自然語言處理技術提高網路入侵偵測的效能和精準度，並彌補傳統入侵檢測系統對動態隱匿性攻擊檢測能力的不足，以提高對新型網路攻擊的適應能力。

三、研究範圍

本研究使用由加拿大網絡安全研究院於 2017 年釋出的 CIC-IDS2017 資料集。這個資料集被廣泛應用於評估入侵檢測系統的性能，旨在為研究模擬現實世界中網路攻擊和正常流量情境的數據。CIC-IDS2017 提供包括良性背景流量以及各類網路攻擊（如阻斷服務攻擊、暴力破解、惡意程式等）在內的真實網路流量，並從上述網路封包中提取 80 多個可用於訓練入侵檢測系統的流量特徵，且對每筆流量預先標記了攻擊手法或正常流量供研究人員使用。

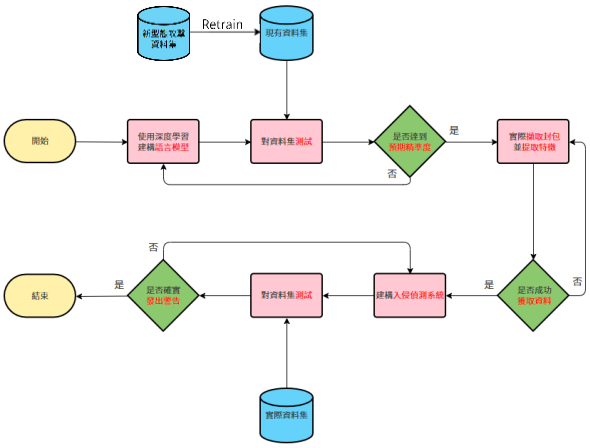
資料集真實反映了現實世界中背景流量不成比例地多於惡意流量的情況。因此我們將正常和惡意流量的比例調整為一致，最後將用於實驗的數據隨機排序，以確保分布平衡，提高模型的準確性和可靠性。

四、使用技術方法

本研究利用 Python 的 SimpleTransformers 套件 [4] 建構一個用於分析流量的自然語言模型。SimpleTransformers 支援多種任務和預訓練自然語言模型，並具備相較於傳統 Transformers 架構更快速訓練及更便利部署的優勢。我們在相同的參數設置下，評估 BERT、RoBERTa、ALBERT 等模型對 CIC-IDS2017 資料集的表現，並根據混淆矩陣、精確率、召回率、ROC 曲線及 F1 score 等指標，選擇最適合的模型。目標是確保模型在 CIC-IDS2017 資料集上的 F1 score 能達到 0.95 以上。

此外，我們使用了增量學習 (Incremental Learning) 技術，使模型能夠持續學習新的攻擊特徵，並應對不斷演變的攻擊手法。

五、架構流程



圖一：研究流程圖

研究流程如圖一，主要分為五步驟：

(一) 建構高效能的流量分析語言模型
(二) 利用 TShark 定期監控並捕獲網路封包
(三) 使用 CICFlowMeter 對封包進行處理和特徵提取
(四) 將特徵與語言模型結合進行分析
(五) 實作入侵檢測系統（IDS）視窗，發出警告通知

六、工具說明

(一) SimpleTransformers - 訓練及預測封包的預訓練模型
(二) TShark - 捕獲並解析網路封包
(三) CICFlowMeter - 從 pcap 檔案提取網路流量特徵
(四) CustomTkinter – 實作 IDS 自定義介面之套件
(五) Metasploit – 測試實際攻擊虛擬機是否能夠正常偵測

七、實驗結果

在成功訓練語言模型並達到 F1 score 0.99 的表現後，我們進一步進行了系統的實驗性測試。在實驗環境中，我們架設了兩台虛擬機，一台用作攻擊機並安裝了 Metasploit，負責發動網路攻擊；另一台則作為靶機，用於接收並處理來自攻擊的流量。靶機配置了 TShark 工具，用來攔截並擷取封包數據，將其回傳至我們的主機進行特徵擷取與分析。

在模型進行特徵擷取並分析封包內容後，我們的系統能夠有效地辨識出惡意流量，並及時向管理員發出警報。這次的實驗不僅驗證了我們模型的精確度與可靠性，還讓我們更有信心在日後的研究中進一步提升模型的效能與應用範疇。

八、參考文獻等

[1] Zoho Corporation Pvt. Ltd., “Common cyberattacks to look out for”, 2023. [Online]. Available: https://www.manageengine.com/log-management/cyber-security-attacks/common-types-of-cyber-attacks.html. [Accessed: 15 5 2024].

[2] A. Gharib, I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "An Evaluation Framework for Intrusion Detection Dataset," 2016 International Conference on Information Science and Security (ICISS), Pattaya, Thailand, pp. 1-6, 2016, doi: 10.1109/ICISSEC.2016.7885840.

[3] H. Kozushko, “Intrusion Detection: Host-Based and Network-Based Intrusion Detection Systems,” vol. 11, 2003.

[4] T. Rajapakse, “About - Simple Transformers,” 2020. [Online]. Available: https://simpletransformers.ai/about/. [Accessed: May 15, 2024].