基於大型語言模型的資料擴增技術在網路威脅情資的攻擊手法分類研究

陳羿琪, 陳怡安, 李熹琳, 黃意婷

國立臺灣科技大學

前言

- 網路威脅情資(Cyber Threat Intelligence，CTI)主要用於提供已知威脅與攻擊事件的資訊，其中包含入侵指標(Indicators of Compromise，IoCs)，以及其攻擊相關描述等等資訊。
- MITRE ATT&CK 框架
  - 描述惡意程式或攻擊組織採用的攻擊的生命週期，被視為開源的 CTI 資料來源，該框架將相關資料以攻擊策略(Tactic)、攻擊手法(Technique)與流程(Procedure)的方式呈現。

研究挑戰	處理方法
目前研究大多使用 MITRE ATT&CK 框架提供的流程範例，然而攻擊手法的流程範例數量並不平均。	利用大型語言模型進行資料擴增，進行攻擊手法分類器的深度學習訓練。
當給定一篇 CTI 報告，有部分文字描述與攻擊手法無關，不利於直接做攻擊手法分類。	建立攻擊手法相關度二元分類器，以區分 CTI 報告中，給定的文字描述是否與攻擊手法相關。

文獻探討

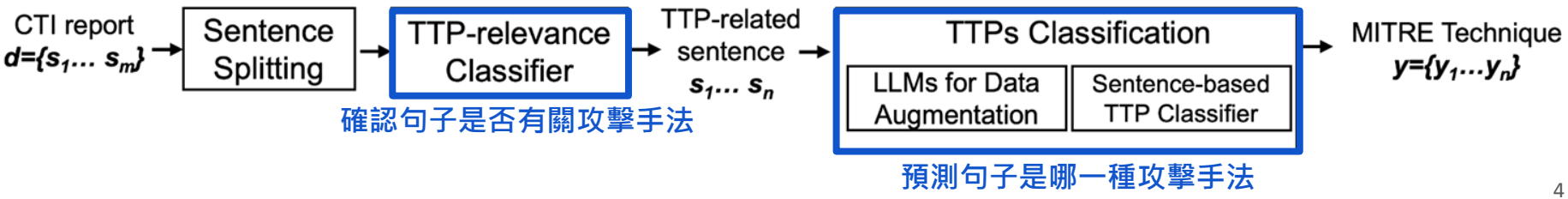
- 多項最新研究致力於自動將 CTI 報告對應到 MITRE ATT&CK 框架中的攻擊手法。

Baseline	Input	Features	Model	Output
AttacKG	Document	攻擊行為圖	計算相似度	Techniques
TTPDrill	Sentence	Subject/Object/Action→BM25 TF-IDF	計算相似度	Tactics and Techniques
LADDER	Sentence	Named entity and relations	計算相似度	Techniques
aCTION	Document	Named entity and relations	計算相似度	Attack pattern
TRAM	Sentence	Embedding	預訓練的 SciBERT	Techniques
rcATT	Document	TF-IDF	線性支援向量機	Tactics and Techniques

問題定義

- 給定 CTI 報告 $\boldsymbol{d} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_m}\}$，將一篇 CTI 報告拆成 m 個句子 s。
  - 我們發展攻擊手法相關度二元分類器，判斷給定的句子是否有關攻擊手法 $\{\boldsymbol{s_1}, \ldots, \boldsymbol{s_n}\}$。
  - 我們發展攻擊手法分類器，判斷每個句子各自對應到 MITRE ATT&CK 框架中哪一個攻擊手法 MITRE Technique (TTP) $\mathbf{y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}\}$。

系統流程圖



攻擊手法相關度二元分類器
TTP-relevance Classifier

- 使用基於 Transformer Encoder 的模型作為本研究的分類器基礎骨幹，用來學習 relevance 相關之知識。
- 我們提取 [CLS] token 來當作分類特徵，用一層線性層搭配 Sigmoid 來作為分類器。
- 目標：判斷一個句子是否描述攻擊手法。
- 這個分類任務的類型是 binary classification。

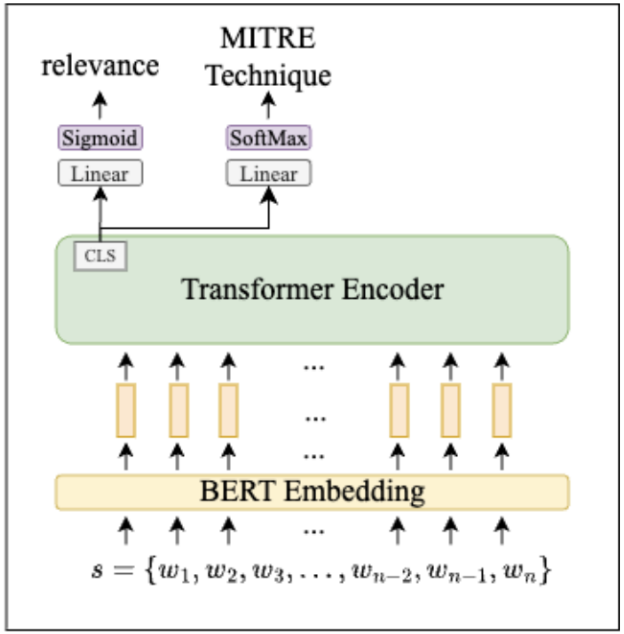
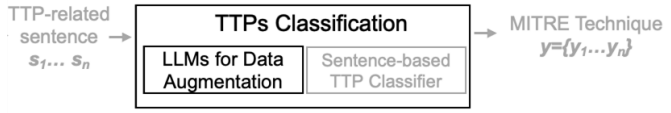
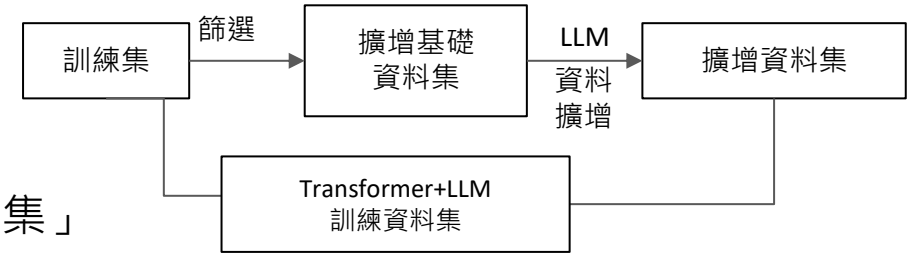


圖 2: 分類器骨幹模型架構圖，分別用於攻擊手法相關度二元分類器與攻擊手法分類器

資料擴增(1/4)





只針對攻擊手法分類器進行資料擴增：

1)篩選適合用於資料擴增的「擴增基礎資料集」

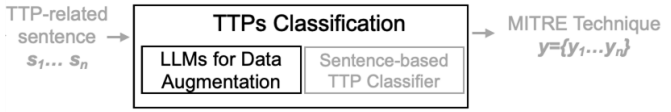
- 從訓練資料集中剔除符合規則的內容，如指令、API 函式與字數過短。

2)使用兩種提示詞進行資料擴增

- 對「擴增基礎資料集」的每一筆資料，使用生成提示詞透過改寫句子生成新資料，
- 使用驗證提示詞，驗證擴增資料的標籤正確性。如果不是，則重新生成。

3)合併「擴增資料集」和原本的訓練集，整理為「Transformer + LLM 訓練資料集」，用於訓練 Sentence-based TTP Classifier。

資料擴增(2/4)：生成提示詞

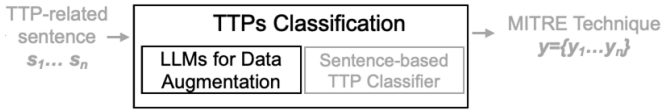


目的是生成新的標籤資料。輸入一個句子及其攻擊手法標籤，讓大型語言模型根據標籤生成 {num_sentences} 個經過改寫的句子。

You are an expert at MITRE ATT&CK framework.
The following sentence is a procedure example of {label}.
{sentence}
Please generate {num_sentences} paraphrased sentences directly without any additional comments or confirmations.

圖 3: 用於生成擴增資料的提示詞

資料擴增(3/4)：驗證提示詞

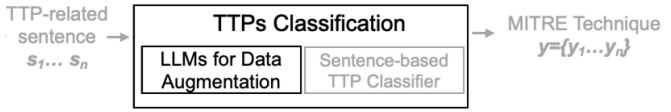


目的是確保生成的 {sentences} 符合給定的攻擊手法 {label}。如果不符合，就重新生成，直到通過驗證才能加入擴增資料集。

You are an expert at MITRE ATT&CK framework.
Please determine if each of the following sentences matches the technique {label}.
Respond 'yes' if it matches and 'no' if it does not.
Here are the sentences: {sentences}

圖 4: 用於驗證擴增資料的提示詞

資料擴增(4/4)：擴增資料範例

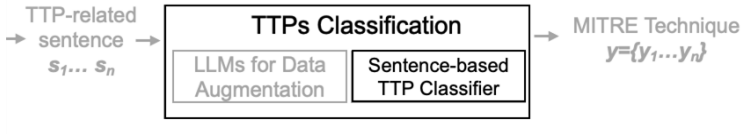


1. 同義詞替換，例如將 added junk bytes 替換為 incorporated irrelevant bytes。
2. 更改語句結構，例如將句尾的 over HTTP 改寫為 In its HTTP command and control activity 並調換至句首。
3. 由於大型語言模型具有基礎電腦科學知識，因此能將 C2 改寫為 command and control。
4. 參考標籤內容，能為整個句子加上有關攻擊手法意涵的總結，如 as a form of obfuscation。

Procedure Example (T1001.001 Data Obfuscation: Junk Data)	
Original	SUNBURST added junk bytes to its C2 over HTTP.
Rephrased	In its HTTP command and control activity, SUNBURST incorporated irrelevant bytes as a form of obfuscation.
TRAM (T1041 Exfiltration Over C2 Channel)	
Original	which will then send the results to the C2 server
Rephrased	After executing this procedure, the outcomes will be sent to the command and control server.

圖 5: 擴增資料範例

攻擊手法分類器
Sentence-based TTP Classifier



- 沿用 TTP-relevance Classifier 使用基於 Transformer Encoder 的模型作為基礎。
- 目標：判斷一個句子描述哪個攻擊手法。
- 這個分類任務的類型是 single-label classification，也就是我們所訓練的模型將可以用於將一個具有攻擊手法意涵的句子，分類至一個其所描述的攻擊手法。

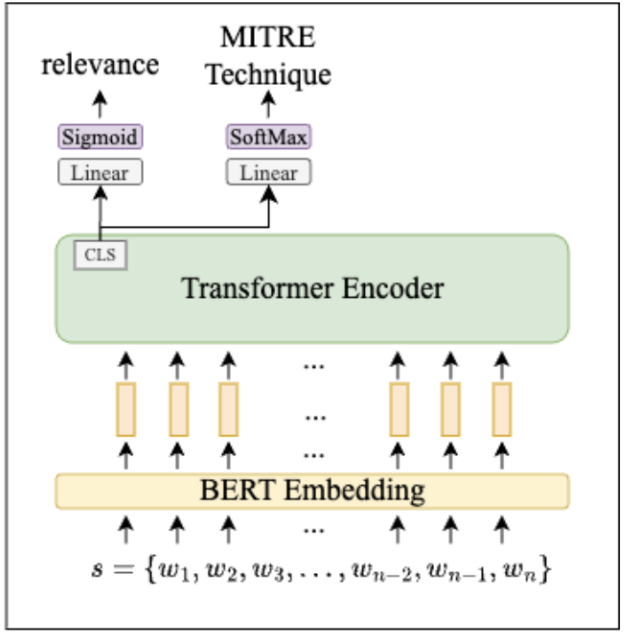


圖 2: 分類器骨幹模型架構圖，分別用於攻擊手法相關度二元分類器與攻擊手法分類器

實驗問題

- **RQ1:**

  本研究所提出的系統架構在攻擊手法分類上的準確性如何？

- **RQ2:**

  當給定一篇 CTI 報告，本研究提出的系統如何幫助攻擊手法上的分類？

資料集：攻擊手法相關度實驗資料集

- 攻擊手法相關度資料需要包含與攻擊手法相關和不相關的句子
- 使用 MITRE TRAM (Threat Report ATT&CK Mapper)專案中的 multi_label.json 資料集，其包含與攻擊手法無關的句子。
- 此資料集以 1/0 表示文字描述是否有對應的攻擊手法。
- 該資料以 8:1:1 的比例切割成訓練集、驗證集與測試集。
- 我們所建立的二元分類器，精確率及召回率分別達到 73.48% 和 79.98%。

	與攻擊手法無關的句子數(0)	**與攻擊手法有關的句子數(1)**
訓練集	10,836	3,178
驗證集	1,355	397
測試集	1,354	398

資料集：攻擊手法分類實驗資料集

- MITRE 流程範例資料集
  - 資料來源為 MITRE ATT&CK 框架第 14 版的流程範例，並且限定為 Windows 平台的資訊。
  - 只採納有 30 筆以上資料的攻擊手法標籤資料。
- TRAM (single) 資料集
  - 源自於 MITRE 發表的 TRAM 專案。
  - 我們移除以下資料
    - 相同文字描述包含多種攻擊手法
    - 源自於不同來源的相同文字描述
    - 僅有大小寫不同的文字描述
- 依照各攻擊手法標籤佔 8:1:1 的比例，切分為訓練集、驗證集及測試集。

	MITRE 流程範例資料集	**TRAM (single) 資料集**
資料數	11,054	4,754
攻擊手法標籤種類	83	50

比較對象

Baseline	**比較目的**	**描述**
AttcaKG	最新 TTP 分類研究	將 CTI 報告整合成「攻擊行為圖」並與攻擊手法比較
TTPDrill	最新 TTP 分類研究	提取句子中的主詞、受詞、動詞，並與攻擊手法比較
LADDER	最新 TTP 分類研究	提取句子中實體之間的關係，並與攻擊手法比較
TRAM(SciBERT)	最新 TTP 分類研究	使用預訓練模型 SciBERT，重新訓練一個分類器
Transformer + Oversampling	資料擴增方法	按照各類別資料筆數決定生成數量，用以資料擴增
Transformer + WordNet	資料擴增方法	將句子中特定數量的詞語更換同義詞，用以資料擴增
Transformer	資料擴增方法	不做資料擴增，訓練時僅使用原始訓練資料
Transformer + LLM (Our)	我們的資料擴增	使用原始訓練資料加上我們的擴增資料，以訓練模型

RQ1: 攻擊手法分類準確性

- 從 MITRE 資料集當中，透過大型語言模型資料擴增的方法，可以加強分類結果。
- 在 TRAM 資料集的召回率表現略為損失，可能是因為大型語言模型中幻覺的發生，造成模型分類有誤。

表 1: 模型在不同資料集上的表現

Dataset	MITRE 攻擊流程範例			TRAM		
Model	Precision	Recall	F1	Precision	Recall	F1
AttcaKG	2.03%	5.25%	2.13%	2.82%	6.48%	3.77%
TTPDrill	15.77%	34.57%	19.15%	12.95%	36.27%	18.79%
LADDER	59.64%	41.42%	43.71%	58.94%	39.24%	40.38%
TRAM(SciBERT)	83.98%	81.82%	81.47%	77.57%	**76.76%**	75.98%
Transformer	82.48%	78.34%	79.15%	78.32%	75.80%	75.65%
Transformer + Oversampling	82.11%	79.93%	79.81%	74.39%	76.21%	75.30%
Transformer + WordNet	84.25%	78.07%	79.17%	78.14%	76.33%	75.83%
Transformer + LLM (Our)	**85.07%**	**82.58%**	**82.66%**	**79.16%**	75.87%	**76.86%**

RQ2: 案例分析(1/2)

此處選擇使用實驗結果 F1 分數表現最佳的「使用 MITRE 流程範例資料集所訓練的 Transformer + LLM」作為攻擊手法分類器。

1. 在報告中擷取一段有寫出攻擊手法 ID 的段落
2. 為了驗證模型的準確性，將文章中寫出的攻擊手法 ID 作為該句的正確答案，並將攻擊手法 ID 從文字中剔除
3. 將經過處理的段落斷句後，放入系統中幫助句子找到其對應的攻擊手法

含有Technique句子的文章段落

A month after, on February 23rd 2022, ESET Research reported a new Wiper being used against hundreds of Ukrainian systems. The wiper receives its name from the stolen certificate (T1588.003) it was using to bypass security controls “Hermetica Digital Ltd” (T1588.003). According to a Reuters article, the certificate could have also been obtained by impersonating the company and requesting a certificate from scratch. The attackers have been seen using several methods to distribute the wiper through the domain, like: domain Group Policy Object (GPO) (T1484.001), Impacket or SMB (T1021.002) and WMI (T1047) with an additional worm component named HermeticWizard.

The wiper component first installs the payload as a service (T1569.002) under C:\Windows\system32\Drivers\. Afterwards, the service corrupts the first 512 bytes of the MBR of all the Physical Drives, and then enumerates their partitions. Before attempting to overwrite as much data as the wiper can it will delete key files in the partition, like MFT, $Bitmap, $LogFile, the NTUSER registry hive (T1112) and the event logs (T1070.001).

資料清理+斷句

- A month after, on February 23rd 2022, ESET Research reported a new Wiper being used against hundreds of Ukrainian systems.
- The wiper receives its name from the stolen certificate it was using to bypass security controls “Hermetica Digital Ltd” .
- According to a Reuters article, the certificate could have also been obtained by impersonating the company and requesting a certificate from scratch.
- The attackers have been seen using several methods to distribute the wiper through the domain, like: domain Group Policy Object (GPO) , Impacket or SMB and WMI with an additional worm component named HermeticWizard.
- The wiper component first installs the payload as a service under C:\\Windows\\system32\\Drivers\\.
- Afterwards, the service corrupts the first 512 bytes of the MBR of all the Physical Drives, and then enumerates their partitions.
- Before attempting to overwrite as much data as the wiper can it will delete key files in the partition, like MFT, $Bitmap, $LogFile, the NTUSER registry hive and the event logs .

RQ2: 案例分析(2/2)

A month after, on February 23rd 2022, ESET Research reported a new Wiper being used against hundreds of Ukrainian systems. → ***no technique***

T1588 : Obtain Capabilities

The wiper receives its name from the stolen certificate it was using to bypass security controls "Hermetica Digital Ltd" → no technique

According to a Reuters article, the certificate could have also been obtained by impersonating the company and requesting a certificate from scratch. → ***no technique***

T1484 : Domain or Tenant Policy Modification

The attackers have been seen using several methods to distribute the wiper through the domain, like: domain Group Policy Object (GPO), Impacket or SMB and WMI with an additional worm component named HermeticWizard. → T1069

T1047 : Windows Management Instrumentation

T1569 : System Services

The wiper component first installs the payload as a service under C:\Windows\system32\Drivers\. → T1543

Afterwards, the service corrupts the first 512 bytes of the MBR of all the Physical Drives, and then enumerates their partitions. → T1083

T1070 : Indicator Removal

Before attempting to overwrite as much data as the wiper can it will delete key files in the partition, like MFT, $Bitmap, $LogFile, the NTUSER registry hive and the event logs. → ***T1070***

實際 CTI 報告結果。藍色字代表預測結果，紅色字代表正確結果。

結論

主要貢獻：

1. 本研究建立一個以 Transformer Encoder 為骨幹的模型，給定一個句子，模型將辨識該句子是否描述一個攻擊手法，並找出其對應的攻擊手法。
2. 應用大型語言模型進行資料擴增，以增加新的標籤資料。
3. 分析 CTI 報告，基於 MITRE ATT&CK 框架，將文章內容惡意活動對應到 MITRE ATT&CK 框架記錄的攻擊手法。

未來的研究方向：

1. 大型語言模型可能產生「幻覺」：可使用 RAG(Retrieval Augmented Generation，擷取增強生成)擷取 CTI 報告中的資訊來優化大型語言模型，讓大型語言模型生成更加準確或完整的描述。
2. 攻擊手法分類實為多標籤和多分類問題：改進模型架構或採用新的演算法。
3. 提取其他有助於訓練分類器的威脅情資：透過大型語言模型進行命名實體識別，自動化提取 CTI 報告中的命名實體與其他命名實體間的關係，提供更多有力的威脅情資。

Q&A

Thank you!