이종 그래프 트랜스포머(HGT)를 활용한 APT 공격 시나리오 자동 분류

(Automatic Classification of APT Attack Scenarios using Heterogeneous Graph Transformer(HGT))

최준호*
(Jun Ho Choi)

요 약

APT 공격은 정보 탈취, 시스템 파괴, 금전적 이득 등 다양한 목적을 가지며, 각 시나리오에 따라 공격 패턴이 구조적으로 상이하다. 기존 연구는 동종 그래프 기반 접근으로 공격의 이종적 특성을 충분히 반영하지 못하며, 시나리오 자동 분류 연구는 거의 없는 실정이다. 본 연구는 APT 보고서로부터 이종 지식그래프를 자동 구축하고, 이종 그래프 트랜스포머(HGT)를 활용하여 시나리오를 자동 분류하는 방법론을 제안한다. SecureBERT 기반 개체 추출과 규칙 기반 관계 추출을 통해 이종 지식그래프를 구축하였으며, HGT의 meta-relation 기반 어텐션으로 시나리오별 구조적 패턴을 학습하였다. 6,127개 캠페인 데이터셋에서 Accuracy 91.4%, Macro-F1 87.2%를 달성하여 기존 GNN 모델(GCN, GAT, HAN, R-GCN) 대비 Macro-F1에서 최대 10.7% 우수한 성능을 보였다.

■ 중심어 : APT 공격 ; Heterogeneous Graph Transformer ; 공격 시나리오 분류 ; 지식그래프

Abstract

APT attacks pursue diverse objectives such as data espionage, system destruction, and financial gain, exhibiting structurally distinct attack patterns for each scenario. Existing research based on homogeneous graphs fails to adequately capture the heterogeneous nature of attacks, and studies on automatic scenario classification are nearly absent. This study proposes a methodology that automatically constructs heterogeneous knowledge graphs from APT reports and classifies scenarios using Heterogeneous Graph Transformer (HGT). Heterogeneous knowledge graphs were built through SecureBERT-based entity extraction and rule-based relation extraction, and scenario-specific structural patterns were learned using HGT's meta-relation-based attention mechanism. The proposed method achieved 91.4% accuracy and 87.2% macro-F1 on 6,127 campaigns, outperforming GCN, GAT, HAN, and R-GCN by up to 10.7%.

■ keywords : APT Attack ; Heterogeneous Graph Transformer ; Attack Scenario Classification ; Knowledge Graph

Ⅰ. 서 론

APT(Advanced Persistent Threat) 공격은 특정 조직이나 국가를 대상으로 장기간에 걸쳐 은밀하게 수행되는 지능형 사이버 공격으로, 2010년 Stuxnet 사건 이후 전 세계적으로 심각한 위협이 되고 있다. 최근 APT 공격은 단순한 정보 탈취를 넘어 랜섬웨어를 통한 금전 탈취, 공급망 침투를 통한 대규모 피해 확산, 암호화폐 채굴을 위한 자원 탈취 등 목적이 다양화되고 있다. 공격 기법 또한 고도화되어 제로데이 취약점 악용, 신뢰받는 소프트웨어 업데이트 프로세스 악용, 다단계 Kill Chain을 통한 은밀한 침투가 증가하고 있으며, 단일 공격이 복수의 목적을 동시에 추구하는 복합 공격 형태로 진화하고 있다. 이러

* 정회원, 조선대학교 자유전공학부
이 논문은 2025학년도 조선대학교 학술연구비의 지원을 받아 연구되었음(K206934009).

접수일자 : 2026년 01월 05일
수정일자 : 2026년 02월 09일
게재확정일 : 2026년 02월 11일
교신저자 : 최준호 e-mail : xdman@chosun.ac.kr

한 공격 목적과 전술의 다양화에도 불구하고 공격 시나리오를 자동으로 분류하고 대응 전략을 수립하는 연구는 여전히 부족한 실정이다. APT 공격은 각 목적에 따라 공격 패턴이 구조적으로 상이하다. 본 연구에서 APT 공격 시나리오는 공격의 최종 목적에 따라 구분되는 공격 유형을 의미하며, 동일한 시나리오 내 공격들은 사용하는 악성코드, 공격 기법, 표적 자산 등에서 유사한 패턴을 보인다. 공격 시나리오를 정확히 식별하면 공격자의 의도를 파악하고 피해 범위를 예측할 수 있어, 효과적인 대응이 가능하다.

기존 APT 탐지 연구는 주로 규칙 기반 시그니처 매칭이나 이상 탐지에 의존하였다. 그러나 새로운 공격 변종 대응이 어렵고 공격의 최종 목적 파악에 한계가 있다. 최근 GNN(Graph Neural Network)을 활용한 연구들이 공격 시퀀스나 시스템 프로비넌스 그래프를 모델링하여 탐지 성능을 개선하였으나[1,2], 대부분 동종 그래프를 가정하여 공격 개체와 관계의 타입 정보를 구분하지 않는다. 이는 malware, technique, tool 등 개체 타입과 uses, targets, exploits 등 관계 타입 정보를 손실시켜, 랜섬웨어의 파일 암호화 패턴과 정보 탈취의 자격증명 유출 패턴을 구조적으로 구분할 수 없다. 또한 공격 탐지나 단계 예측에 집중하며, 시나리오 자동 분류 연구는 거의 없는 실정이다.

본 논문은 APT 공격 보고서로부터 이종 지식그래프(Heterogeneous Knowledge Graph)를 자동 구축하고, HGT(Heterogeneous Graph Transformer)를 활용하여 공격의 최종 목적에 따른 시나리오를 자동으로 분류하는 방법론을 제안한다. APT 공격의 이종적 특성을 명시적으로 모델링하기 위해 SecureBERT 기반 개체 추출과 규칙 기반 관계 추출을 통해 9개 노드 타입과 9개 엣지 타입으로 구성된 이종 지식그래프를 구축한다. 구축된 그래프는 HGT의 meta-relation 기반 어텐션을 통해 시나리오별 구조적 패턴을 학습하여 공격 목적을 판별한다.

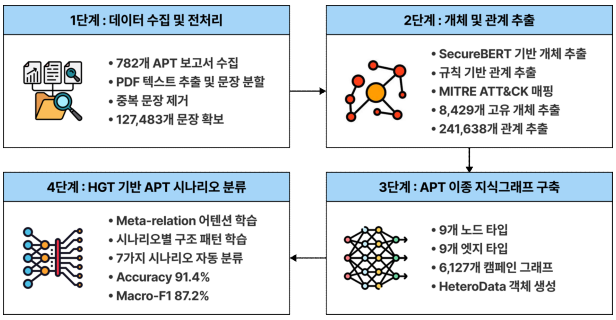

그림 1. 본 논문의 연구 프로세스

본 논문의 구성은 다음과 같다. 2장에서는 GNN 기반 APT 분석 연구와 HGT 관련 연구를 검토한다. 3장에서는 SecureBERT 기반 APT 이종 지식그래프 구축 과정을 설명하고, 4장에서는 HGT 기반 시나리오 자동 분류 방법을 제시한다. 5장에서는 실험 결과 및 분석을 기술하고, 6장에서 결론 및 향후 연구 방향을 제시한다.

II. 관련 연구

1. GNN 기반 APT 공격 패턴 분석

GNN은 노드와 엣지로 구성된 그래프 구조 데이터를 학습하는 딥러닝 모델로, GCN(Graph Convolutional Network)은 이웃 정보 집계를 통한 기본 구조를 제시하였고, GAT(Graph Attention Network)는 어텐션 메커니즘으로 이웃 노드의 중요도를 차별적으로 학습한다[3]. 사이버보안 분야에서는 악성코드 탐지를 위한 함수 호출 그래프 학습, 네트워크 침입 탐지를 위한 트래픽 그래프 모델링 등 다양한 연구가 진행되었다[4]. APT 공격 분석에서는 공격 시퀀스를 프로비넌스 그래프(Provenance Graph), 즉 시스템 이벤트 간 인과관계를 추적하는 그래프로 표현하여 공격 경로를 추론하는 연구와 Kill Chain 단계를 그래프로 모델링하여 공격 단계를 예측하는 연구가 수행되었다[5]. 그러나 기존 GNN 기반 APT 연구는 대부분 동종 그래프를 가정하여 모든 노드와 엣지를 동일한 타입으로 취급한다. 이는 악성코드, 공격 기법, 도구 등 다양한 개

체 타입과 사용, 표적화, 악용 등 다양한 관계 타입으로 구성된 APT 공격의 이종적 특성을 충분히 반영하지 못한다. 또한 대부분의 연구는 공격 탐지나 단계 예측에 집중하며, 공격의 최종 목적에 따른 시나리오 분류에 대한 연구가 미비하다.

2. 이종 그래프 신경망 및 HGT

이종 그래프는 여러 타입의 노드와 엣지가 혼합된 그래프로, 각 노드와 엣지가 서로 다른 의미와 속성을 가질 수 있어 실세계의 복잡하고 다층적인 관계를 표현하는 데 적합하다[6,7]. 초기 이종 그래프 신경망인 HAN(Heterogeneous Attention Network)은 meta-path 기반 어텐션을 제안하였으나 사전 정의된 meta-path에 의존하며, R-GCN(Relational Graph Convolutional Network)은 엣지 타입별 가중치를 학습하지만, 노드 타입 정보는 활용하지 못한다[8,9]. HGT는 meta-relation 기반 어텐션 메커니즘을 핵심으로 한다. Meta-relation은 <source_type, edge_type, target_type> 트리플로 정의되며, HGT는 각 meta-relation마다 독립적인 어텐션 가중치를 학습하여 동일한 연결이라도 노드 타입 조합에 따라 다른 중요도를 부여한다[10,11]. 예를 들어 <User, writes, Paper>와 <User, cites, Paper> 관계는 서로 다른 의미를 가지며, HGT는 이를 구분하여 학습한다. HGT는 학술 논문 네트워크, 추천 시스템 등 다양한 도메인에서 우수한 성능을 보였으며, 사이버보안 분야에서는 위협 인텔리전스 분석과 취약점 예측에 적용되었다.

III. SecureBERT 기반 APT 시나리오 지식그래프 구축

1. 데이터 수집 및 전처리

HGT 모델 학습을 위한 이종 그래프 데이터셋 구축을 위해 GitHub 저장소[1]의 공개 APT 보고서를 활용한다. 이 저장소는 Mandiant, CrowdStrike, Kaspersky 등 주요 보안업체의 2010년부터 2024년까지 위협 분석 보고서를 APT 그룹별로 정리하고 있다. 수집된 847개 보고서 중 명확한 APT 캠페인을 다루고 3페이지 이상의 기술 분석을 포함하는 782개를 선별하였다. PDF 텍스트 추출은 PyPDF2와 pdfplumber를 사용하고, 추출된 텍스트는 header/footer 제거, 섹션 분류, 문장 분할 과정을 거친다. 중복 문장은 수식 1의 Jaccard 유사도로 제거한다.

$$J(s_1, s_2) = \frac{|W(s_1) \cap W(s_2)|}{|W(s_1) \cup W(s_2)|} \quad (1)$$

여기서 W_s는 문장 s의 단어 집합이며, $J(s_1, s_2) \geq 0.9$인 경우 중복으로 제거한다. 이 임계값은 예비 실험에서 0.9 미만 시 의미적으로 다른 문장이 과도하게 제거되고, 0.9 초과 시 중복 제거 효과가 미미함을 확인하여 결정하였다. 최종적으로 127,483개 문장을 추출하였으며, 보고서 당 평균 163.1개 문장을 포함한다.

2. APT 이종 지식그래프 설계

전통적인 GCN, GAT 등은 모든 노드와 엣지를 동일한 타입으로 취급하는 동종 그래프를 입력으로 받는다. 이러한 구조에서는 악성코드와 자격증명이 동일한 노드로, 사용과 표적화가 동일한 엣지로 처리되어 의미적 차이가 손실된다. 그러나 APT 공격은 악성코드, 공격 기법, 도구, 표적 자산 등 서로 다른 역할을 가진 요소들과 사용, 표적화, 악용 등 서로 다른 의미의 관계로 구성된 이종 그래프 형태이다. 동종 그래프로 표현하면 랜섬웨어의 파일 암호화 패턴과 정보 탈취의 자격증명 유출 패턴을 구조적으로 구분할 수 없다. 이종 그래프는 서로 다른 타입의 노드와 엣지가 혼재된 그래프로, G = (V, E, A, R)로 정의된다. V는 노드 집합, E는 엣지 집합, A는 노

[1] https://github.com/blackorbird/APT_REPORT

드 타입 집합, R은 엣지 타입 집합이다. 본 연구는 APT 공격을 이종 그래프로 모델링하여 HGT가 타입 조합별로 차별화된 학습을 수행할 수 있게 한다. 표 1은 9가지 노드 타입을 보여준다.

표 1. 노드 타입 정의

노드 타입	설명	예시
CampaignAPT	APT 캠페인	WannaCry_2017, Petya_2017, APT28_Ukraine_2022
Malware	악성코드 및 악성 소프트웨어	ransomware, trojan, backdoor, worm
Technique	공격 기법 및 전술	phishing, lateral_movement, privilege_escalation
Tool	공격 도구 및 유틸리티	mimikatz, psexec, powershell, cobalt_strike
Vulnerability	취약점 개념	zero_day, buffer_overflow, sql_injection
Infrastructure	공격 인프라	command_and_control, proxy, botnet, domain
Target_Asset	표적 자산 및 데이터	credential, email, database, financial_data
System_Resource	시스템 자원	registry, process, memory, file_system
Attack_Action	공격 행위	encrypt, exfiltrate, dump, hijack, inject

캠페인 타입은 개별 APT 공격 사례를 나타내며, 해당 캠페인에 포함된 모든 개체를 연결하는 중심 역할을 한다. 나머지 8가지 타입은 공격 구성 개체로, 여러 캠페인에 걸쳐 재사용된다. 엣지 타입은 표 2와 같이 정의된다.

표 2. 엣지 타입 정의

엣지 타입	Source→Target	의미
CONTAINS	Campaign→ 모든 개체 타입	캠페인이 포함하는 개체
PERFORMS	Attack_Action→Technique	행위가 수행하는 기법
TARGETS	Attack_Action→Target_Asset	행위가 표적하는 자산
USES	Technique/Malware→Tool	기법/악성코드가 사용하는 도구
EXPLOITS	Technique→Vulnerability	기법이 악용하는 취약점
ACCESSES	Malware/Tool→System_Resource	악성코드/도구가 접근하는 자원
EXTRACTS	System_Resource→Target_Asset	자원에서 추출하는 자산
UTILIZES	Technique/Malware→Infrastructure	기법/악성코드가 활용하는 인프라
LEADS_TO	Technique→Technique	시간적 선후 관계

HGT의 meta-relation은 <source_type, edge_type, target_type> 트리플로 정의되며, 본 연구에서는 노드 타입과 엣지 타입 조합에 따라 23개의 유효한 meta-relation이 존재한다.

예를 들어 <Attack_Action, PERFORMS, Technique>, <Malware, USES, Tool>, <Technique, LEADS_TO, Technique> 등이다. HGT는 각 meta-relation 별로 독립적인 어텐션 가중치를 학습하므로, 같은 USES 엣지라도 <Malware, USES, Tool>과 <Technique, USES, Tool>은 서로 다른 중요도로 학습된다. 이를 통해 악성코드가 도구를 사용하는 패턴과 공격 기법이 도구를 사용하는 패턴을 구조적으로 구분할 수 있다.

3. APT 이종 지식그래프 생성

본 논문에서는 통합 APT 지식그래프를 구축한 후 캠페인별 서브그래프를 추출한다. 이를 통해 개체 노드 재사용으로 효율적인 표현 학습과 공통 패턴 포착이 가능하다.[12]. APT 보고서 텍스트에서 개체를 추출하기 위해 사이버보안 텍스트로 사전학습된 SecureBERT를 활용한다[13]. SecureBERT는 CVE, 보안 블로그, 위협 리포트 등으로 학습된 BERT 기반 모델로, ‘exploit’, ‘ransomware’, ‘lateral movement’ 등 보안 도메인 특화 용어를 효과적으로 인식한다. 추출된 개체는 품사 태깅과 Sentence-BERT 기반 의미 유사도 계산을 통해 표 1의 9가지 개체 타입으로 분류된다. 동사형 단어는 Attack_Action으로, 시스템 관련 명사는 System_Resource로, 정보 자산 명사는 Target_Asset으로 분류된다. 추출된 개체는 형태소 정규화로 전처리한다. 동사 시제를 통일(encrypted →encrypt)하고 복수형을 단수로 변환(files→file)한 후, Sentence-BERT로 임베딩한다. 개체 간 유사도는 코사인 유사도로 계산한다. 두 개체의 유사도는 수식 2와 같다.

$$sim(c_i, c_j) = \frac{(v_{c_i} \cdot v_{c_j})}{(\|v_{c_i}\| \cdot \|v_{c_j}\|)} \quad (2)$$

여기서 v_c는 개체 c의 임베딩 벡터이다. 유사도가 0.88 이상인 개체를 하나의 클러스터로 통합한다. 예를 들어 ‘steal password’, ‘dump credential’, ‘extract login’은 ‘extract_credential’로 통합된다. 127,483개 문장으로부터 총 98,547개의 개체를 추출하였으며, 정규화 후 8,429개의 고유 개체 노드로 수렴하였다.

개체 노드 간 관계는 규칙 기반 패턴 매칭과 의존 구문 분석으로 추출한다. 53개 명시적 동사 패턴('X

uses Y′, 'X exploits Y′ 등)을 적용하고, 패턴이 없으면 spaCy로 구문 트리 거리 3 이내 개체 쌍의 관계를 추정한다. 시간적 선후 관계는 MITRE ATT&CK Kill Chain 순서로 LEADS_TO 엣지를 생성한다. 신뢰도는 명시적 동사 패턴 0.95, MITRE 매핑 0.85, Kill Chain 순서 0.75, 의존 구문 분석 0.70이며, 0.65 미만은 제거한다.

782개 보고서로부터 6,847개 APT 캠페인을 식별하였으며, 각 캠페인의 Campaign 노드를 생성하여 개체 노드와 CONTAINS 엣지로 연결한다. 명확한 시나리오 판별이 가능한 6,127개(89.5%)를 최종 데이터셋으로 선정하였다. 전체 APT 이종 지식그래프의 통계는 표 3과 같다.

표 3. 전체 APT 이종 지식그래프 통계

항목	수치
총 노드 수	14,556개 (개체 노드 8,429 + Campaign 노드 6,127)
총 엣지 수	481,967개 (개체 간 엣지 241,638 + CONTAINS 엣지 240,329)
캠페인당 평균 개체 수	39.2개
캠페인당 평균 엣지 수	39.4개

HGT의 그래프 수준 분류를 위해 각 Campaign 노드를 중심으로 서브그래프를 추출한다. Campaign 노드 c에 대해 CONTAINS 엣지로 연결된 모든 개체 노드 집합 V_c를 추출하고, V_c 내 모든 노드 쌍 간 존재하는 엣지 집합 E_c를 추출하여 서브그래프 $G_c = (V_c, E_c, A, R)$을 구성한다. Campaign으로부터 6,847개 서브그래프를 추출하였으며, 캠페인당 평균 39.2개 노드와 39.4개 엣지를 포함한다. 각 캠페인 서브그래프에 7가지 시나리오 중 하나의 레이블을 할당한다. 시나리오별 핵심 노드와 보조 노드는 MITRE ATT&CK 프레임워크와 수집된 APT 보고서의 빈도 분석을 결합하여 정의하였다. 예를 들어 Ransomware 시나리오는 MITRE ATT&CK의 "Data Encrypted for Impact(T1486)" Technique을 참조하고, Ransomware 보고서에서 빈도가 높은 encrypt, ransom, inhibit_recovery를 핵심 노드로 지정하였다. 캠페인 c의 시나리오 s에 대한 점수는 수식 3과 같다.

$$S(c,s) = \sum_{i \in C^s_{core}} I(i \in V_c) + 0.5 \sum_{j \in C^s_{support}} I(j \in V_c) \quad (3)$$

여기서 C^s_{core}는 시나리오 s의 핵심 노드 집합, $C^s_{support}$는 보조 노드 집합, V_c는 캠페인 c의 노드 집합이다. 레이블링 결과 6,847개 중 6,127개(89.5%)에 명확한 레이블을 할당하였으며, 시나리오별 분포는 Ransomware 2,247개(36.7%), Data Espionage 1,682개(27.5%), Credential Harvesting 1,023개(16.7%), Cryptojacking 487개(7.9%), Destructive 289개(4.7%), Botnet 254개(4.1%), Supply Chain 145개(2.4%)이다.

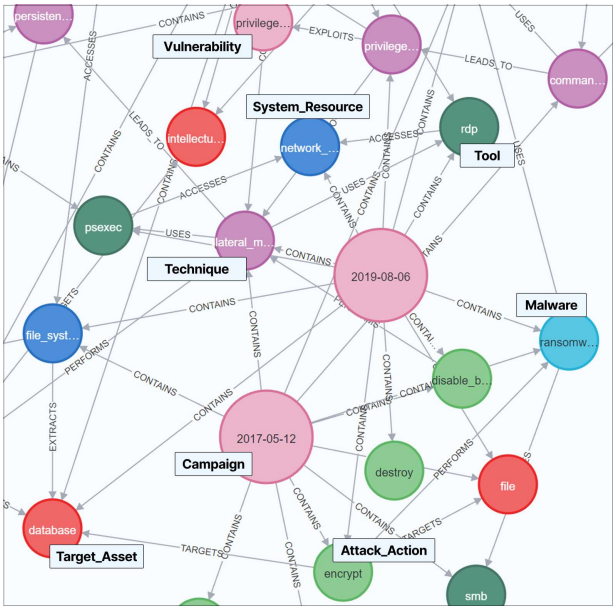

그림 2. 구축된 APT 이종 지식그래프 구조

그림 2는 구축된 APT 이종 지식그래프 구조이다. Campaign 노드가 CONTAINS 엣지로 개체를 연결하며, 개체 노드들은 9개 타입으로 분류되어 TARGETS, USES, LEADS_TO 등 meta-relation으로 연결된다. 각 노드는 Sentence-BERT 768차원 임베딩 벡터를 특징으로 갖는다. 데이터셋은 시간 기반으로 훈련 60%, 검증 20%, 테스트 20%로 분할한다.

Ⅳ. HGT 기반 APT 시나리오 자동 분류

1. 문제 정의 및 분류 파이프라인

본 논문의 목표는 새로운 APT 공격 보고서가 주어졌을 때, 해당 공격이 7가지 시나리오 중 무엇에 해당하는지 자동으로 판별하는 것이다.

표 4. APT 시나리오 자동 분류 파이프라인

입력 : 새로운 APT 보고서 PDF
"Unknown_APT_Campaign_2025.pdf"

↓

1단계 : 보고서 분석 및 이종 그래프 생성
- PyPDF2로 텍스트 추출 및 문장 분할 (NLTK)
- SecureBERT로 개체 추출 및 9개 노드 타입 분류
- 규칙 패턴(53개) + spaCy로 관계 추출 및 9개 엣지 타입 분류
- HeteroData 객체로 이종 그래프 구축

출력 : 캠페인 이종 그래프 Gc
- 42개 노드 (Malware 3개, Attack_Action 5개, ...)
- 44개 엣지 (TARGETS 8개, USES 5개, LEADS_TO 4개, ...)
- 12개 meta-relation 사용
- 노드 특징 : Sentence-BERT 768차원 임베딩

3장 지식그래프 활용 :
- "encrypt", "powershell" 등 → 3장의 8,429개 정규화 노드와 매칭
- <Attack_Action, TARGETS, Target_Asset> 8개 발견
  → Ransomware의 대표 패턴 (3장 학습 결과)

↓

2단계 : HGT 기반 시나리오 분류
[3장에서 학습된 HGT 모델 적용]
- 3개 HGT 층 통과 (meta-relation별 어텐션 계산)
- 타입별 평균 풀링 → 그래프 표현 h_G (2304차원)
- 분류기 → 7개 시나리오 확률 예측

학습된 지식 활용 :
- 3장의 Ransomware 2,247개 학습 패턴 적용
- <Attack_Action, TARGETS, Target_Asset> 어텐션 0.87 (높음)
- 새 그래프에서 이 패턴 8개 → Ransomware로 판별

출력 : 시나리오 확률 분포
[0.968, 0.021, 0.008, 0.002, 0.001, 0.000, 0.000]

↓

3단계 : 분류 결과 출력 및 해석
최종 결과 :
✓ Ransomware Campaign ····· 96.8%
Destructive Attack ····· 2.1%
Data Espionage ····· 0.8%
기타 ····· 0.3%

분류 근거 :
- <Attack_Action, TARGETS, Target_Asset> 관계 8개 (핵심 패턴)
- encrypt, file, ransomware 노드 포함 (Ransomware 특징)
- 3장 Ransomware 평균(32.1 노드)과 유사한 크기(42 노드)

APT 시나리오 자동 분류 문제는 매핑 함수 $f: G_c \rightarrow \{s_1, s_2, \cdots, s_7\}$로 정의되며, G_c는 캠페인 이종 그래프, s_i는 i번째 시나리오이다. 전체 파이프라인은 표 4와 같다. 1단계는 3장과 동일하게 그래프를 생성하고, 2단계에서 HGT로 시나리오를 판별한다.

2. HGT 기반 시나리오 판별 메커니즘

HGT는 meta-relation 기반 어텐션을 통해 시나리오별 구조적 패턴을 학습한다. <source_type, edge_type, target_type> 트리플마다 독립적인 어텐션 가중치를 학습하여, 시나리오별로 중요한 관계를 차별적으로 포착한다. HGT는 L개 층으로 구성되며, 각 층에서 노드 표현을 업데이트한다. 노드 타입 τ(v)에 따라 다른 변환 행렬이 적용되고, meta-relation별 어텐션이 수식 4와 같이 계산된다.

$$h_v^{(l)} = \sigma\Big(\sum_{u \in N(v)} \alpha(u,v) \cdot W_M^{\phi(e)} \cdot V_u^{(l-1)}\Big) \quad (4)$$
$$\alpha(u,v) = softmax\big(W_A^{\phi(e)} \cdot [Q_v \| K_u]\big)$$

여기서, $h_v^{(l)}$은 l번째 층 노드 v의 임베딩, $N(v)$는 이웃 노드 집합, $\alpha(u,v)$는 어텐션 가중치, $W_M^{\phi(e)}$는 메시지 변환 행렬이다. L개 층을 거친 후 타입별 평균 풀링으로 그래프 표현을 생성하고, 분류기를 통해 수식 5와 같이 시나리오 확률을 예측한다.

$$h_G = \bigoplus_{\tau \in A} \Big(\frac{1}{|V_\tau|} \sum_{v \in V_\tau} h_v^{(L)}\Big) \quad (5)$$
$$\hat{y} = softmax\big(W_{classifier} \cdot h_G + b\big)$$

여기서, h_G는 그래프 표현 벡터, A는 노드 타입 집합, V_τ는 타입 t의 노드 집합, $W_{classifier}$는 분류기 가중치이다. HGT가 학습한 APT 시나리오별 핵심 패턴은 표 5와 같다.

표 5. APT 시나리오별 학습된 핵심 패턴

시나리오	핵심 Meta-relation	구조적 특징
Ransomware Campaign	<Attack_Action, TARGETS, Target_Asset>	평균 노드 32.1개, 밀집도 높음
Data Espionage	<Technique, LEADS_TO, Technique>	평균 5.7 hops, 긴 시간적 체인
Supply Chain Attack	<Malware, USES, Tool>	USES 관계 비율 27.3%
Credential Harvesting	<System_Resource, EXTRACTS, Target_Asset>	memory→credential 패턴
Destructive Attack	<Attack_Action, ACCESSES, System_Resource>	시스템 자원 직접 접근
Cryptojacking	<Attack_Action, ACCESSES, System_Resource>	CPU/GPU 자원 강조
Botnet Recruitment	<Malware, UTILIZES, Infrastructure>	C2 통신 강조

예를 들어, 새로운 보고서 "Malware encrypts corporate files and demands Bitcoin"에서 생성된 그래프가 (encrypt, TARGETS, file), (encrypt, TARGETS, database) 관계를 다수 포함하면, HGT는 <Attack_Action, TARGETS, Target_Asset> meta-relation의 어텐션 가중치를 높게 설정하여 Ransomware로 분류한다. 분류 결과는 [Ransomware : 96.8%, Destructive : 2.1%, 기타 :

1.1%]와 같은 결과처럼 확률 분포로 출력된다.

3. HGT 모델 학습 전략 및 평가 설계

HGT 모델의 학습은 3장에서 구축한 6,127개의 레이블링된 캠페인 그래프를 사용한다. 데이터는 시간 기반으로 훈련 60%(3,676개), 검증 20%(1,226개), 테스트 20%(1,225개)로 분할한다. 클래스 불균형을 고려하여 수식 6과 같이 가중 교차 엔트로피 손실을 사용한다.

$$L = -\frac{1}{N}\sum_{c=1}^{7} w_c \cdot \sum_{i=1}^{N} y_{i,c} \cdot \log(\hat{y}_{i,c}) \qquad (6)$$

$$w_c = \frac{N_{total}}{C \cdot N_c}$$

여기서 N은 배치 크기, $y_{i,c}$는 샘플 i의 실제 레이블(클래스 c이면 1, 아니면 0), $\hat{y}_{i,c}$는 예측 확률이다. w_c는 클래스 c의 가중치로, 샘플이 적은 클래스에 더 높은 가중치를 부여하여 클래스 불균형을 보정한다. 예를 들어, Ransomware(2,247개)는 낮은 가중치를, Supply Chain(145개)은 높은 가중치를 갖는다. 모델 구조 및 학습 설정은 표 6과 같다.

표 6. HGT 모델 설정 및 베이스라인

항목	HGT	GCN	GAT	HAN	R-GCN
그래프 타입	이종	동종	동종	이종	이종
층 수	3	3	3	3	3
숨겨진 차원	256	256	256	256	256
Attention 헤드	8	–	8	8	–
Meta-relation 어텐션	○	×	×	부분	×

학습은 Adam 옵티마이저(학습률 0.001, weight decay 0.0001)를 사용하며, 배치 크기는 32이다. 소수 클래스(Supply Chain, Botnet)는 2배 오버샘플링하고, 조기 종료는 검증 손실이 20 epochs 동안 개선되지 않으면 적용한다. 평가 지표는 Accuracy, Macro-F1, Weighted-F1을 사용하며, 시나리오별 Precision, Recall, F1-score와 혼동 행렬을 분석한다. 베이스라인 모델(GCN, GAT, HAN, R-GCN)과 비교하여 HGT의 meta-relation 기반 어텐션의 효과를 검증한다. 동종 그래프 모델(GCN, GAT)은 모든 노드와 엣지를 동일 타입으로 통합하여 입력하므로 타입 정보 손실이 발생한다.

Ⅴ. 실험 및 평가

1. 실험 환경 및 데이터셋

본 논문은 6,127개의 레이블링된 APT 캠페인 그래프를 이용하여 실험을 수행하였다. 모든 모델은 NVIDIA GeForce RTX 3070 GPU에서 PyTorch 2.0을 사용하여 학습하였다. 베이스라인 모델로는 동종 그래프 기반 GCN, GAT와 이종 그래프 기반 HAN, R-GCN을 사용하였으며, 모든 모델은 3개 층, 256차원의 임베딩으로 통일하였다. 평가 지표로는 Accuracy, Macro-F1, Weighted-F1을 사용하였으며, 5회 반복 실험의 평균±표준편차로 결과를 제시하였다.

2. 전체 분류 성능 비교

HGT는 Accuracy 91.4%, Macro-F1 87.2%, Weighted-F1 90.1%로 모든 베이스라인을 크게 상회하였다. 특히 Macro-F1에서 차선 모델(R-GCN) 대비 7.1%p 향상되어, 소수 클래스에서도 우수한 성능을 보임을 확인하였다. 동종 그래프 모델(GCN, GAT)은 노드와 엣지 타입 정보를 손실하여 상대적으로 낮은 성능을 보였으며, 이종 그래프 모델(HAN, R-GCN) 중에서는 관계별 가중치를 학습하는 R-GCN이 meta-path 기반 HAN보다 우수하였다. HGT는 meta-relation별 독립적인 어텐션을 통해 시나리오별 구조적 패턴을 효과적으로 학습하였다. 특히 개체 및 관계 추출 과정의 오류에도 불구하고 HGT가 높은 분류 성능(91.4%)을 달성한 것은 의미가 있다. 이는 HGT가 개별 노드/엣지가 아닌 그래프 수준의 구조적 특징(예: <Attack_Action, TARGETS, Target_Asset> meta-relation의 빈도와 분포)을 학습하므로, 일부 개체나 관계 오류가 전체 시나리오 판별에 미치는 영향이 제한적이기 때문이다. 표 7은 제안 모델(HGT)과 베이스라인 모델들의 분류 성능을 측정한 결과이다.

표 7. 베이스라인 모델 대비 전체 분류 성능

모델	그래프 타입	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
GCN	동종	78.3 ± 1.2	68.4 ± 1.8	76.1 ± 1.4
GAT	동종	81.5 ± 0.9	72.6 ± 1.5	79.4 ± 1.1
HAN	이종	85.2 ± 0.8	78.3 ± 1.2	83.7 ± 0.9
R-GCN	이종	86.7 ± 0.7	80.1 ± 1.1	85.2 ± 0.8
HGT(제안모델)	이종	91.4 ± 0.5	87.2 ± 0.9	90.1 ± 0.6

그림 3은 HGT의 학습 곡선을 보여준다. 훈련 손실과 검증 손실이 안정적으로 수렴하며, 약 80 epoch에서 최적 성능에 도달하였다. 과적합 없이 학습이 진행되었음을 확인할 수 있다.

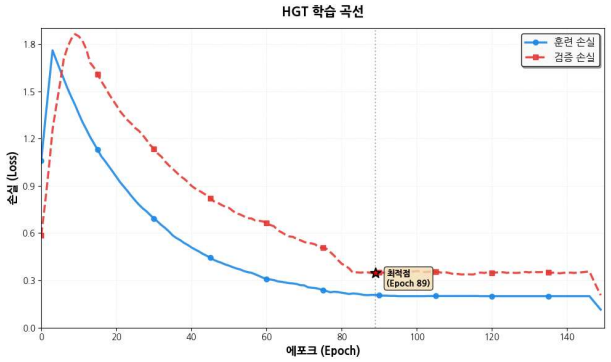

그림 3. HGT 학습 곡선

3. APT 시나리오별 성능 분석

표 8은 HGT의 APT 시나리오별 세부 성능을 측정한 결과이다. 다수 클래스인 Ransomware(F1 97.8%)와 Data Espionage(F1 94.5%)는 높은 성능을 보였다.

표 8. APT 시나리오별 분류 성능(HGT)

시나리오	샘플 수	Precision (%)	Recall (%)	F1-score (%)
Ransomware Campaign	449	97.5	98.1	97.8
Data Espionage	337	94.2	94.8	94.5
Credential Harvesting	205	91.5	90.5	91.0
Cryptojacking	97	87.6	86.4	87.0
Destructive Attack	58	84.8	83.2	84.0
Botnet Recruitment	51	81.5	80.5	81.0
Supply Chain Attack	29	76.2	74.0	75.1
평균 (Macro)	–	87.6	86.8	87.2
평균 (Weighted)	–	91.8	91.5	91.6

이는 두 시나리오가 <Attack_Action, TARGETS, Target_Asset>과 <Technique, LEADS_TO, Technique> 같은 명확한 구조적 패턴을 갖기 때문이다. Credential Harvesting(F1 91.0%)과 Cryptojacking(F1 87.0%)도 우수한 성능을 달성하였다. 소수 클래스인 Supply Chain Attack(F1 75.1%)은 상대적으로 낮은 성능을 보였다. 주요 원인은 전체 데이터에서 Supply Chain이 145개(2.4%)로 희소하고, 테스트 셋에서도 29개로 표본 수가 적어 변별 패턴 학습이 어렵기 때문이고, 고유 패턴의 변별력 부족이 영향을 미친다. 향후 데이터 증강과 대조 학습을 통해 개선할 계획이다.

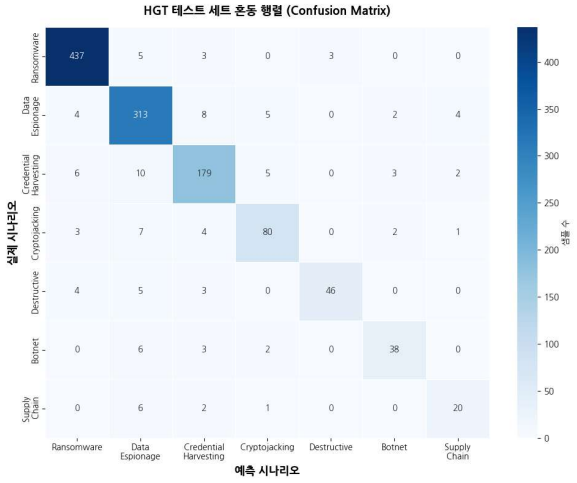

그림 4. HGT 테스트 세트 혼동 행렬

그림 4는 HGT의 테스트 세트 혼동 행렬을 보여준다. 대부분 시나리오가 대각선에 집중되어 올바르게 분류되었으나, Supply Chain Attack이 Data Espionage로 일부 오분류되었다(29개 중 6개, 20.7%). 이는 두 시나리오가 긴 Kill Chain과 은밀한 침투 패턴을 공유해 그래프 구조가 유사하게 나타날 수 있음을 시사한다. 본 실험을 통해 제안한 HGT 기반 APT 시나리오 분류 방법이 기존 GNN 모델 대비 우수한 성능을 보임을 확인하였다. 특히 이종 그래프 구조와 meta-relation 기반 어텐션이 시나리오별 구조적 패턴을 효과적으로 학습하여, 다양한 APT 공격 시나리오를 높은 정확도로 자동 분류할 수 있음을 입증하였다.

Ⅵ. 결론 및 향후 연구

본 논문은 APT 공격 보고서로부터 이종 지식그래프를 자동 구축하고, Heterogeneous Graph Transformer(HGT)를 활용하여 7가지 APT 시나리오를 자동 분류하는 방법론을 제안하였다. SecureBERT 기반 개체 추출과 규칙 기반 관계 추

출을 통해 9개 노드 타입과 9개 엣지 타입으로 구성된 이종 지식그래프를 구축하였으며, 6,127개의 레이블링된 캠페인 그래프로 HGT 모델을 학습하였다. 제안 방법은 APT 공격의 이종적 특성을 명시적으로 모델링하여 공격 개체와 관계의 타입 정보를 보존하였으며, meta-relation 기반 어텐션을 통해 시나리오별 구조적 패턴을 차별적으로 학습하였다. 실험 결과, 제안 방법은 테스트 세트에서 Accuracy 91.4%, Macro-F1 87.2%를 달성하여 기존 GNN 모델 대비 최대 10.7%p 향상된 성능을 보였다. Ransomware Campaign(F1 97.8%)과 Data Espionage(F1 94.5%)에서 높은 분류 성능을 달성하였으며, 소수 클래스인 Supply Chain Attack도 F1 75.1%의 실용적인 성능을 보였다. 본 연구는 개체 및 관계 추출 오류가 분류 성능에 영향을 줄 수 있는 한계가 있다. 규칙 기반 접근은 53개 패턴으로 제한되어 암묵적 관계를 누락할 수 있으며, SecureBERT도 일부 개체 경계 모호성 문제를 가진다. 향후 연구에서는 다중 레이블 분류로 확장하여 복합 목적 공격을 다루고, End-to-End 학습 구조를 도입하여 개체/관계 추출과 시나리오 분류를 동시에 최적화함으로써 오류 전파를 줄일 계획이다. 또한 실시간 APT 탐지 시스템과 통합하여 SOC 환경에서 자동 분류 결과를 기반으로 대응 전략을 제시하는 실무 시스템 구축을 진행할 예정이다.

REFERENCES

[1] M. Schlichtkrull, et al., “Modeling relational data with graph convolutional networks,” *In European semantic web conference, Springer*, pp. 593-607, 2018.

[2] W. Ren, et al., “APT attack detection based on graph convolutional neural networks,” *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 184, 2023.

[3] C. Do Xuan and M. H. Dao, “A novel approach for APT attack detection based on combined deep learning model,” *Neural Computing and Applications*, vol. 33, no. 20, pp. 13251-13264, 2021.

[4] M. Jiang, et al., “An Approach for APT Attack Scenario Construction Based on Dynamic Attack Graphs,” *GLOBECOM 2024-2024 IEEE Global Communications Conference, IEEE*, pp. 3087-3092, Dec. 2024.

[5] C. Zhang, et al., “CNN-KOA-BiGRU: A high-accuracy APT detection model based on deep learning networks,” *In 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE*, pp. 2342-2347, 2024.

[6] Q. Wang, et al., “HANDOM: Heterogeneous attention network model for malicious domain detection,” *Computers & Security*, vol. 125, p. 103059, 2023.

[7] A. G. Vrahatis, et al., “Graph attention networks: a comprehensive review of methods and applications,” *Future Internet*, vol. 16, no. 9, p.318, 2024.

[8] R. F. T. Ceskoutsé, et al., “HeteroKGRep: Heterogeneous Knowledge Graph based Drug Repositioning,” *Knowledge-Based Systems*, vol. 305, p. 112638, 2024.

[9] Z. Hu, et al., “Heterogeneous graph transformer,” *Proc. of the web conference 2020*, pp. 2704-2710, 2020.

[10] K. Liu, et al., “A review of knowledge graph application scenarios in cyber security,” arXiv preprint, arXiv:2204.04769, Apr. 2022.

[11] Y. Ren, et al., “CSKG4APT: A cybersecurity knowledge graph for advanced persistent threat organization attribution,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 5695-5709, 2022.

[12] 이승주, 안석호, 이의종, 서영덕, “이종 데이터 간 관계 모델링을 통한 개인화 추천 시스템의 지식 그래프 확장 기법,” *스마트미디어저널*, 제12권, 제4호, 27-40쪽, 2023년 5월

[13] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” arXiv preprint arXiv:1908.10084, Aug. 2019.

저 자 소 개

최준호(정회원)

2004년 조선대학교 전자계산학과 박사 졸업.

2014년～현재 조선대학교 자유전공학부 부교수.

<주관심분야 : 지식처리, 지식그래프, 자연어처리, 딥러닝, 정보보안>