하이퍼-관계형 지식 그래프 기반 지식 증류를 통한 보이스 피싱 탐지 성능 향상

김성수
한국전자통신연구원
sungsoo@etri.re.kr

Enhancing Voice Phishing Detection Performance via Knowledge Distillation on a Hyper-Relational Knowledge Graph

Sungsoo Kim
Electronics and Telecommunications Research Institute

요 약

본 논문은 보이스피싱 통화 기록을 하이퍼-관계적 지식 그래프로 변환과 이종 그래프기반 모델 경량화 기법을 제안한다. 기존의 단순 텍스트 분류 방식과 달리 제안 기법의 하이퍼-관계적 지식 그래프 기반 표현은 통화 내 등장하는 엔티티와 그 사이의 복잡한 관계를 구조적으로 포착할 수 있으며, 의미론적 맥락과 상황적 특성을 동시에 표현할 수 있는 장점을 갖는다. 또한, 이종 그래프의 구조적 복잡성과 의미론적 풍부한 특성을 고려한 새로운 지식 증류 프레임워크를 제안한다. 실험 결과, 학생 모델은 교사 대비 파라미터 96% 감소와 AUC-ROC 0.9943으로 보이스 피싱 탐지 성능 향상을 보였다.

I. 서론

최근 고도화되는 보이스피싱 범죄는 기존의 규칙 기반 시스템이나 텍스트 분류 모델의 한계를 드러낸다 [1]. 특히, 보이스피싱 통화는 다양한 엔티티와 그들 간의 복잡한 관계가 얽힌 구조적 특성을 갖는다. 그래프 신경망(GNN)을 활용한 보이스피싱 탐지 연구가 주목받고 있지만, GNN은 높은 계산 복잡도로 인해 실시간 탐지 환경에 적용하는 데 제약이 따른다 [2].

복잡한 교사 모델의 지식을 경량의 학생 모델로 이전하는 *지식 증류*(Knowledge Distillation) 접근법이 주목받았다 [3]. 하지만 기존 지식 증류 연구는 동종(homogeneous) 그래프에 초점을 맞추거나 단순한 지식 전달에 국한되어, 이종(*heterogeneous*) 그래프의 복잡한 구조와 풍부한 의미론적 정보를 효과적으로 증류하는 데 한계가 있다 [4].

본 논문은 이종 그래프의 복잡한 구조와 풍부한 의미론적 정보를 효과적으로 증류하기 위한 지식 그래프 모델링 및 지식 증류 기법을 제안한다. 본 논문의 기여는 세가지로 요약할 수 있다.

- *구조화된 의미 표현*: 효과적인 보이스 피싱 탐지를 위한 하이퍼-관계형 지식 그래프 표현과 학습 방법론을 제안한다.
- *이종 그래프 지식 증류*: 이종 그래프 특성을 고려한 다중 손실 함수 기반 그래프 지식 증류 프레임워크를 제시한다.
- *효과적인 모델 경량화*: 실험을 통해 제안하는 그래프 기반 지식 증류 방법론의 경량화 모델 효율성과 효과성을 제시한다.

II. 하이퍼-관계형 지식 그래프 기반 지식 증류

문제 정의. 보이스 피싱 학습 데이터셋은 $\mathcal{D} = \{(\mathcal{T}_i, y_i)\}_{i=1}^{N}$로 표현하며, 여기서, $\mathcal{T}_i$는 i-번째 통화 기록 텍스트이며, y_i는 보이스피싱 여부('0'- normal, '1' - voice phishing)를 나타낸다 [5].

본 연구는 통화 기록 텍스트($\mathcal{T}$)를 분석하여 보이스피싱 여부를 판별하는 이진 분류 문제로 정의한다.

데이터 변환. 통화 기록 $\mathcal{T}_i$를 *하이퍼-관계형 지식 그래프*(HRKG)로 표현하기 위해 데이터 변환을 진행한다. HRKG $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{Q})$는 노드 집합($\mathcal{V}$), 에지 집합($\mathcal{E}$), 관계 타입 집합($\mathcal{R}$), 한정자(qualifier) 피처 벡터 집합($\mathcal{Q}$)으로 구성된다.

데이터 변환을 위해, KoBERT 기반 NER을 활용하여 통화 기록 텍스트에 대한 엔티티, 관계, 한정자 탐지를 수행한다. 이는 텍스트 내 키워드를 기반으로 복잡한 상호 연결성을 그래프 구조로 표현함으로써, 관계의 세부 맥락을 포착한다.

교사 모델. 제안하는 교사 모델은 이종 그래프 트랜스포머(Heterogeneous Graph Transformer, HGT) 기반으로 어텐션 메커니즘을 통해 노드 간의 중요도를 학습한다. 그림 1과 같이, HRKG를 입력받고, 노드 피처를 선형 레이어를 통해 투영(차원 수: 64)한 후 HGT 컨볼루션 레이어(레이어 수: 3)를 통과한다. 최종적으로 노드 임베딩을 추출하여 분류 작업에 사용한다.

학생 모델. 제안 모델은 그래프 인코더와 텍스트 인코더로 구성된다. 그래프 인코더는 경량화된 GNN으로 그래프 구조 정보를 인코딩하며, 텍스트 인코더는 KoBERT로 텍스트 의미를 인코딩한다. 두 피처 벡터는 *피처 융합*(Feature Fusion) 모듈에서 결합되어 최종 로짓(logits)을 출력한다.

지식 증류. 본 논문의 지식 증류 기법은 *Teacher-to-Student 증류* 기법에 속한다 [3]. 그림 1의 프레임워크는 HRKG의 복잡한 구조를 유지하면서 학생 모델의 효율성을 극대화한다. 이는 *표현 증류*(Representation distillation)와 *로짓 기반 증류*(Logit-based distillation)로 구현된다. 표현 증류는 학생이 교사의 특정 중간 레이어 지식을 전달받고, 로짓 기반 증류는 출력 로짓을 모방한다.

손실 함수. 기본적으로 교사 모델과 학생 모델 간의 지식 증류 손실 함수 $\mathcal{L}_{\text{kd}}$은 다음과 같이 정의할 수 있다.

$$\mathcal{L}_{\text{kd}} = D_{\text{KL}}(\mathcal{K}_T, \mathcal{K}_S), \tag{1}$$

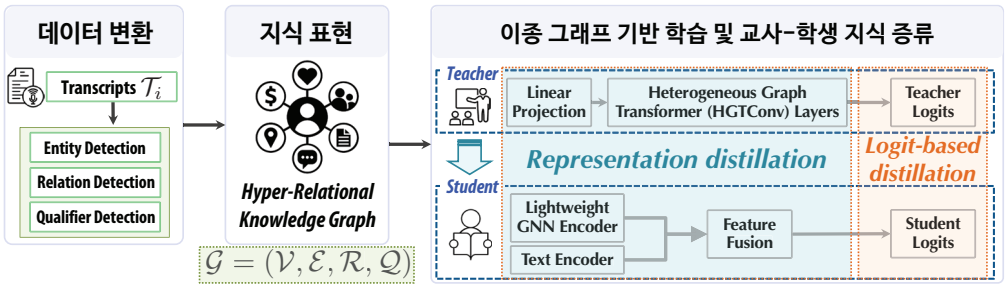

그림 1: 보이스 피싱 탐지를 위한 하이퍼-관계형 지식 그래프 기반 지식 증류 프레임워크

여기서 D_{KL}는 이진 분류에 적합한 KL-발산(Kullback–Leibler Divergence)을 나타내며, $\mathcal{K}_T$와 $\mathcal{K}_S$는 각각 교사 및 학생 모델로부터 얻은 지식이다. 본 논문에서는 *지식 증류 손실* D_{KL}, *이진 교차 엔트로피 손실* $\mathcal{L}_{\mathrm{BCE}}$, 그래프 라플라시안(Laplacian) 기반 *위상 보존 손실* $\mathcal{L}_{\mathrm{Topo}}$를 결합한 다중 손실 함수를 적용했다.

$$\mathcal{L}_{\mathrm{kd}} = \alpha \cdot D_{\mathrm{KL}}\left(\sigma\left(\frac{\mathbf{z}_S}{T}\right) \middle\| \sigma\left(\frac{\mathbf{z}_T}{T}\right)\right) + (1-\alpha)\cdot \mathcal{L}_{\mathrm{BCE}} + \beta \cdot \mathcal{L}_{\mathrm{Topo}} \quad (2)$$

여기서, $\mathbf{z}_T$, $\mathbf{z}_S$는 출력 로짓, T는 temperature, α, β는 가중치다. 제안 방식은 학생 모델이 교사 모델의 지식을 효율적으로 흡수하면서도 필요한 최소한의 파라미터만을 사용하도록 보장한다.

III. 실험 및 결과

실험 설정. 실험은 Apple M3 Max CPU, 128GB RAM, Python 3.12, PyTorch Geometric 2.2 환경에서 진행했다. 학습 데이터는 KorCCVi v2 데이터셋 [5]을 사용했다. 이 데이터셋은 695개의 보이스 피싱 샘플과 2,232개의 비-보이스 피싱 샘플을 포함하여 총 2,927개로 구성되어 있으며, 훈련 70%, 검증 15%, 테스트 15%로 사용했다. 또한, 제안 모델의 평가지표는 정확도, 정밀도, 재현율, F1-스코어, AUC-ROC 점수를 활용했다.

특성	교사 모델	학생 모델	비율
파라미터수	970,979	30,081	96% (감소)
추론시간 (ms)	12.1996	0.2635	46배 (향상)
모델크기 (MB)	3.7040	0.1147	32배 (압축)

표 1: 교사모델과 학생모델의 특성 비교

표 1과 같이, 학생 모델은 파라미터 96% 감소, 추론 시간 46배 단축을 달성했다. 그림 2와 같이, 학생 모델은 교사 모델보다 모든 지표에서 우수한 성능을 보였다. 특히, 보이스피싱 탐지에서 가장 중요한 지표인 auc_roc 점수에서 교사 모델의 0.9754를 능가하는 0.9943을 기록하며 압도적인 탐지 성능을 보였다. 이 결과는 본 논문에서 제안한 지식 증류 기법이 모델 경량성과 탐지 성능을 동시에 확보하는 효과적인 해결책임을 시사한다.

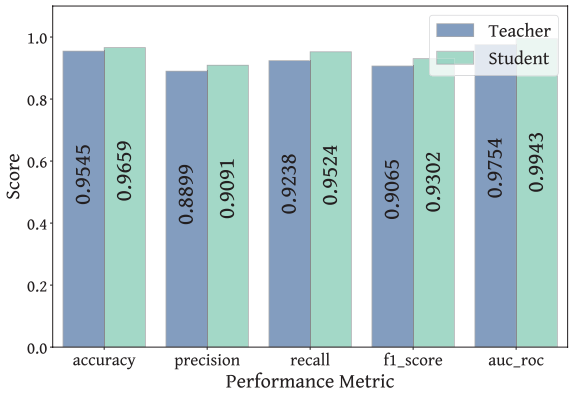

그림 2: 그래프 기반 지식 증류 모델 성능 비교

IV. 결론

본 논문은 효율적인 보이스피싱 탐지를 위해 하이퍼-관계형 지식 그래프(HRKG) 기반 지식 증류 프레임워크를 제안했다. 보이스피싱 통화의 복잡한 구조적 특성을 HRKG로 모델링하고, 이종 그래프 기반 교사 모델의 지식을 경량화된 학생 모델로 전달하는 다중 손실 함수 기반 지식 증류 기법을 제시했다. 실험을 통해 학생 모델이 교사 모델 대비 우수한 성능과 효율성을 달성함을 입증했다. 향후 온디바이스 모델의 설명 가능성(XAI)과 신종 보이스피싱에 대한 지속 학습 방안을 탐구할 계획이다.

Acknowledgement. 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2025-02215393, (2세부) 알려지지 않은 신종 보이스피싱 탐지 · 예측 기술개발)

참고 문헌

[1] S. Kim, "Efficient Voice Phishing Detection using the Agentic AI Approach," in *2025 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pp. 1–6, IEEE, 2025.

[2] J. Fu, C. Li, Z. Zhao, and Q. Zeng, "Heterogeneous graph knowledge distillation neural network incorporating multiple relations and cross-semantic interactions," *Inf. Sci.*, vol. 658, p. 120004, 2024.

[3] Y. Tian, S. Pei, X. Zhang, C. Zhang, and N. V. Chawla, "Knowledge Distillation on Graphs: A Survey," *ACM Comput. Surv.*, vol. 57, no. 8, pp. 189:1–189:16, 2025.

[4] G. Sun, X. Zhang, J. Ni, and D. Song, "Towards Heterogeneous Continual Graph Learning via Meta-knowledge Distillation," *CoRR*, vol. abs/2505.17458, 2025.

[5] M. K. M. Boussougou, P. Hamandawana, and D. Park, "Enhancing Voice Phishing Detection Using Multilingual Back-Translation and SMOTE: An Empirical Study," *IEEE Access*, vol. 13, pp. 37946–37965, 2025.