JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 2. pp. 369-378, Feb. 2023

Check for updates

효율적인 이상 탐지를 위한 적대적 도메인 적응 기법

황 현 정[1] · 김 강 석[2*]
[1]아주대학교 지식정보공학과 석사과정
[2*]아주대학교 사이버보안학과 교수

Adversarial Domain Adaptation Technique for Efficient Anomaly Detection

Hyun-Jung Hwang[1] · Kangseok Kim[2*]

[1]Master's Course, Department of Knowledge Information Engineering, Ajou University, Suwon 16499, Korea
[2*]Professor, Department of Cyber Security, Ajou University, Suwon 16499, Korea

[요 약]

본 논문에서는 효율적인 이상 탐지를 위한 적대적 도메인 적응 기법을 제안한다. 제안된 방법은 생성적 적대 네트워크(GAN)에 사용되는 적대적 학습을 통해 두 도메인 사이의 확률적 데이터 분포 차이를 최소화한다. 제안된 모델은 분류 성능 평가 지표를 사용하여, 제안된 도메인 적응 모델의 이상 탐지 성능을 도메인 적응을 적용하지 않은 모델과 비교하여 평가한다. 실험결과 제안된 모델이 비교 모델보다 학습시간을 67% 단축하고 유사한 성능을 보이는 효율적인 이상 탐지 성능을 보였으며, 기존 전이 학습(Transfer Learning)의 학습 과정과 달리 도메인 적응 방법을 통해 타겟 데이터가 소스 데이터 보다 상대적으로 큰 경우에도 효과적임을 확인하였다.

[Abstract]

In this paper, we propose adversarial domain adaptation techniques for efficient anomaly detection. The proposed method minimizes the difference in stochastic data distribution between the two domains through adversarial learning used in generative adversarial networks (GANs). We evaluate the anomaly detection performance of the proposed domain adaptation model by using classification performance evaluation indicators compared to models without domain adaptation. Experiments show that the proposed model reduces learning time by 67% compared to comparative models, shows similar performance, and shows efficient anomaly detection performance, and unlike the learning process of transfer learning, it is effective even when target data is relatively larger than source data.

색인어 : 정보 보안, 딥러닝, 적대적 도메인 적응, 이상 탐지, 시계열 데이터

Keyword : Information Security, Deep Learning, Adversarial Domain Adaptation, Anomaly Detection, Time Series Data

http://dx.doi.org/10.9728/dcs.2023.24.2.369

Received 15 December 2022; **Revised** 10 January 2023
Accepted 11 January 2022

***Corresponding Author; Kangseok Kim**

Tel: +82-31-219-2496
E-mail: kangskim@ajou.ac.kr

Ⅰ. 서 론

현재 우리 사회는 4차 산업혁명 후 다양한 디지털 기기의 사용에 따른 인터넷 사용량이 급증함에 따라 디지털 생태계가 확장되면서 로그4j(Log4j), 랜섬웨어(Ransomware) 등과 같은 사이버 보안의 위협이 기하급수적으로 증가하고 있다. 특히 인터넷으로 가상화된 IT리소스를 서비스로 제공하는 클라우드 서비스(Cloud Service)가 증가함에 따라 리눅스(Linux) 환경을 노리는 익스플로잇(Exploit)과 맬웨어(Malware)공격 또한 증가하고 있는 추세이다. 이러한 사이버 공격에 대응하는 방법 중 하나인 침입 탐지 시스템(Intrusion Detection System)은 시스템 공격, 오용, 침입을 방지하는 역할을 하며, 지식 기반 침입 탐지인 오용 탐지(Misuse Detection)와 행위 기반 침입 탐지인 이상 탐지(Anomaly Detection)로 나뉜다. 오용 탐지는 오탐률이 낮지만 새로운 침입 유형에 대한 탐지가 어렵다는 단점이 있고, 이상 탐지[1]는 정상 패턴과 다른 비정상적인 패턴을 식별하는데 사용되는 기술로 새로운 침입 유형의 탐지가 가능하다. 최근 대용량 데이터를 학습하는 인공지능 구현 기술인 기계 학습 및 딥러닝 기술이 발전하면서 이와 결합된 이상 탐지 기술에 대한 연구가 증가하고 있으며, 많은 보안 분야 연구자들은 기계 학습과 딥러닝을 활용하여 기존 방법보다 효율적으로 이상 탐지를 할 수 있는 다양한 방법의 연구를 시도하고 있는 중이다.

기존 이상 탐지 모델은 어느 한 도메인의 정상과 이상(Anomaly) 데이터를 구분할 수 있도록 학습되어 모델과 도메인은 일대일 대응을 이룬다. 그로 인해, 새로운 도메인의 이상 탐지 모델을 구축하기 위해서는 새로운 도메인의 정상과 이상 데이터를 구분할 수 있도록 이상 탐지 모델을 다시 학습시켜야 하기 때문에 학습 시간이 오래 걸리게 되고 실시간 이상 탐지를 어렵게 할 수 있는 한계점이 존재한다. 따라서 본 논문에서는 효율적인 이상 탐지를 위해 도메인 적응(Domain Adaptation) 기법을 적용한 이상 탐지 모델을 제안한다. 도메인 적응 기법은 기존 도메인을 소스 도메인(Source Domain), 새로운 도메인을 타겟 도메인(Target Domain)이라 하며, 학습을 통해 소스 도메인과 타겟 도메인의 확률적 데이터 분포의 차이를 최소화하여 새로운 도메인의 이상 탐지 시간을 단축시키는 방법이다. 자율주행 및 의학 등 많은 분야에서 연구되어온 이상 탐지 모델들은 이미지와 같은 비정형 데이터 영역에서 우수한 성능을 보이며 연구되고 있지만, 정형 데이터 영역에서는 오히려 기계 학습을 활용한 XGBoost(Extreme Gradient Boosting)[2], LightGBM(Light Gradient Boosting Machine)[3] 같은 트리 기반 앙상블(Ensemble)[4]모델이 딥러닝 모델보다 우수한 성능을 보이며 현재까지 사용되고 있다[5].

따라서 본 논문에서는 구조화된 보안 데이터와 도메인 적응 기법의 결합을 통해 이상 탐지에 효율적인 모델을 제안하고자 하며, 제안 모델을 위해 사용한 기법으로 자기지도학습(Self-Supervised Learning)[6] 중 하나인 Word2Vec [7], 시계열 데이터에 적합한 장단기 메모리(LSTM; Long Short-Term Memory)[8], 도메인 적응(Domain Adaptation)[9]이 있다. 제안 모델은 실험 데이터로 ADFA-LD(Australian Defence Force Academy Linux Dataset)[10]와 NGIDS-DS(Next-Generation Intrusion Detection System Dataset)[11] 데이터를 사용하여 실험을 진행하였으며, 리눅스 시스템의 시스템 콜 시퀀스에 적합한 임베딩 기법을 적용하여 데이터 전처리를 수행하였다. 또한, 도메인 적응 모델 중 적대적 도메인 적응(Adversarial Discriminative Domain Adaptation)[12] 모델의 인코더(Encoder)에 LSTM 기반의 인코더를 적용하여 이상 탐지 기법의 개발을 진행하였다.

Ⅱ. 관 련 연 구

이 장에서는 기존의 이상 탐지 방법과 적대적 학습 및 도메인 적응 방법에 대해 간략히 설명한다.

2-1 이상 탐지(Anomaly Detection)

이상 탐지는 데이터에서 예상 동작과 일치하지 않는 패턴을 찾는 문제를 말한다. 이상 탐지는 컴퓨터 시스템 상에 침입을 탐지하는 사이버 침입 탐지(Cyber-Intrusion Detection), 보험, 신용, 금융 관련 데이터에서 불법 행위를 검출하는 사기 탐지(Fraud Detection), 악성 코드를 검출 해내는 악성코드 탐지(Malware Detection), 의학 데이터에 대한 이상치를 탐지하는 의학 이상 탐지(Medical Anomaly Detection) 등 최근 다양한 분야에서 사용되고 연구되고 있다[1].

[13]은 정상 데이터만을 대상으로 딥러닝 모델을 학습하여 이상 침입을 탐지하는 비지도 학습 기반 네트워크 이상 탐지 방법으로 학습이 완료된 오토인코더의 재구성 손실 및 각 레이어 출력에 대한 마할라노비스 거리(Mahalanobis distance)들을 가중합하여 탐지 대상 데이터의 이상치 점수를 도출하고, 도출한 이상치 점수에 임계값을 적용함으로써 네트워크 이상 침입을 효율적으로 탐지하는 모델을 제안하여, 오토인코더의 재구성 손실(Reconstruction loss) 만을 사용하는 경우에 비해 우수한 성능을 보임을 확인하였다.

[14]는 이상 데이터 없이 이미지의 알려지지 않은 이상 패턴을 감지하는 결함 탐지를 위한 고성능 모델을 구축하는 것을 목표로 정상 훈련 데이터만을 사용하는 이상 탐지 모델 CutPaste를 제안하였으며, ImageNet에서 사전 학습된 특징을 전이 하여 AUC Score 96.6을 달성하였다.

2-2 적대적 학습(Adversarial Learning)

기계학습 모델은 복잡하고, 종종 예측을 수행하는 방법을 제대로 이해하지 못할 수 있으며 이러한 문제들은 공격자가

악용할 수 있는 문제점일 수 있다. 모델을 속여 잘못된 예측을 하거나 민감한 정보를 제공받을 수 있고, 가짜 데이터를 사용하여 우리가 모르는 사이에 모델을 손상시킬 수 있다. 적대적 공격에는 모델을 훈련하는 데 사용되는 데이터에 초점을 맞춘 중독 공격(Poisoning attacks), 모델 자체에 초점을 맞춘 회피 공격(Evasion Attack) 등이 있다[15]. 적대적 학습은 이러한 약점들을 해결하는 것을 목표로 한다.

[16]은 타겟 레이블 집합에 대한 명시적인 가정이 없는 장애 진단을 위한 소스 클래스 별, 대상 인스턴스 별 가중치 메커니즘을 사용하는 하이브리드 접근 방식을 제안하였다. 제안된 방법은 추가 이상값 식별자를 사용하여 제안 방법이 대상 레이블 집합을 알지 못하는 상태에서 공유 상태에 대한 클래스 수준 정렬을 달성하면서 알 수 없는 오류 모드를 자동으로 인식할 수 있으며, 두 개의 데이터셋에 대한 실험을 통해 제안 방법의 유효성을 검증하였다.

[17]은 다중 시점 IRM(Imaginative Reasoning Module)과 TALM(Triple Adversarial Learning Module)로 구성된 UDA person re-ID를 위한 삼중 적대적 학습 및 다중 시점 상상 추론 네트워크(TAL-MIRN)를 제안하였다. TALM은 카메라 분류기와 특징 인코더 간의 적대적 학습, 공동 분포 정렬의 적대적 학습, 분류에 사용되는 두 분류기 간의 차이에 대한 적대적 학습으로 구성되어있으며, 제안 모델을 5개의 데이터셋에 적용하여 비교 실험 결과를 통해 최신 방법 보다 우수함을 확인하였다.

[18]은 소스 데이터를 사용할 수 없다는 문제를 해결하기 위해 적대적 기반의 적응 추론, 대조적 카테고리 와이즈 매칭(Contrastive Category-wise Matching), 자기 지도 회전(Self-Supervised Rotation) 세 가지 구성 요소를 가진 A2Net(Adaptive Adversarial Network)을 제안하였으며, 교차 도메인 벤치마크에 대한 광범위한 실험은 소스 데이터 없이 적응 작업을 해결하는 데 제안된 모델의 효율성을 확인하였다.

2-3 도메인 적응(Domain Adaptation)

도메인 적응은 도메인이 다른 그러나 유사한 새로운 도메인(Target Domain)에 기존 도메인(Source Domain)의 정보를 적응(adaptation) 시켜서 사용하고자 하는 목적을 가지는 연구 분야이다. 기존의 모델이 동작하는 영역을 소스 도메인이라 정의하고 새로운 영역을 타겟 도메인이라 정의하여, 소스 도메인과 타겟 도메인이 같은 클래스끼리는 데이터의 분포가 작아지도록 하고 도메인이 달라도 비슷한 것들은 유사한 특징을 갖도록 하는 방법이다[9].

기계 학습 및 딥러닝 에서는 소스 데이터에서 좋은 성능을 보이는 모델이 타겟 데이터에서 제대로 동작하지 않는 문제가 발생하는데, 이는 소스 데이터에 과하게 적합한 모델의 과대 적합(Overfitting)일 수도 있지만 훈련 데이터와 테스트 데이터의 도메인 이동(Domain Shift) 발생으로 인한 문제일 수 있다. 여기서 도메인 이동은 소스 데이터와 타겟 데이터의 분포 차이를 의미하며, 차이가 심할수록 타겟 데이터의 정확도가 떨어진다.

최근 기존 도메인의 모델과 새로운 도메인의 모델을 동시에 사용하면서 손실 함수를 최소화하는 방향으로 모델 파라미터를 구하는 뉴럴 네트워크 모델 dSNE[19] 방법이 연구되기도 하였다. 도메인 적응은 그라디언트 기반의 적응(Gradient-based Adaptation) 방식과 적대적 학습 기반의 적응(Adversarial-based Adaptation) 방식 두 가지로 볼 수 있다. 그라디언트 기반의 도메인 적응은 타겟 도메인의 라벨이 없어도 학습이 가능하고 역전파(Backpropagation)를 조금 바꾼 단순 구조로 [20]에서 제안하였으며, 적대적 기반의 도메인 적응은 순차적으로 소스 도메인에 대해 먼저 학습한 후, 도메인 적응을 진행하는 방식으로 [12]에서 제안하였다.

Ⅲ. 제 안 방 법 론

본 장에서는 효율적인 이상 탐지를 위해 도메인 적응 기법을 적용하여 서로 다른 두 도메인의 확률적 데이터 분포 차이를 줄이는 모델을 제안하고 설명한다.

3-1 데이터 셋

본 연구에서는 ADFA-LD(Australian Defence Force Academy Linux Dataset)와 NGIDS-DS(Next-Generation Intrusion Detection System Dataset) 데이터셋을 사용하여 실험하였다. 그림 2에서 표기한 ADFA는 본 연구에서 제안하는 도메인 적응 모델의 소스 도메인(Source Domain)으로 사용하였으며, NGIDS는 타겟 도메인(Target Domain)으로서 사용하였다.

1) ADFA-LD 데이터 셋

ADFA 데이터셋은 뉴사우스웨일스 대학교(University of New South Wales) 에서 만든 침입 탐지 데이터셋으로, 리눅스 로컬 서버로부터 수집된 다양한 애플리케이션에서 발생한 가장 최근의 공격 및 취약점에 대한 system call trace로 구성되어 있다[21].

ADFA-LD는 Hydra-FTP, Hydra-SSH, Adduser, JavaMeterpreter, Meterpreter, Webshell 등 6가지 유형의 최신 사이버 공격이 포함된 데이터셋이며 정상 데이터 5,206건, 공격 데이터 746건 총 5,952건의 데이터로 구성되어 있다. [22]에서 기술한 것처럼, KDD 데이터셋은 최신 공격 프로토콜을 반영하지 않고 데이터 손상 및 불일치로 인해 실험에 사용하기 어렵기 때문에 최신 사이버 공격이 포함된 ADFA 데이터셋을 소스 데이터로 사용하였다.

2) NGIDS-DS 데이터 셋

NGIDS 데이터셋은 2017년에 공개된 데이터셋으로 ADFA에서 진행되었던 프로젝트의 일부로, IXIA사의

Perfect Storm이라는 시뮬레이션 장비를 사용해서 생성하였으며, 리눅스 환경에서 수행되는 정상 및 비정상 호스트에서 발생되는 트래픽을 수집한 데이터셋 이다[23].

NGIDS-DS는 정상 데이터 88,791,734건, 공격 데이터 1,262,426건 총 90,054,160건의 데이터를 가지고 있으며 date, time, pro_id 등의 속성과 attack_cat, attack_subcat, label의 속성 타입이 있다. [11]에서 기술한 것처럼 침입 탐지 시스템을 실제 사용하기 전 성능에 대한 현실적인 평가를 얻는 것이 필수적이기 때문에 다양한 기업의 사이버 인프라 수준에서 발생할 수 있는 네트워크 활동에 대해 정상 및 비정상으로 구성된 차세대 침입 탐지 시스템 데이터셋인 NGIDS를 타겟 도메인으로 사용하였다.

3-2 데이터 전처리

1) Data Shuffle

소스 도메인으로 사용하는 ADFA는 데이터셋의 정상 데이터와 공격 데이터 모두 8:2의 비율로 섞어 훈련 데이터셋과 테스트 데이터셋을 구성하였다. ADFA는 공격 데이터의 경우 세부적인 공격 카테고리를 갖고 있으며, 각 카테고리 별 건수가 다른 점을 고려하여 서브 카테고리 별 건수도 8:2로 나누어 데이터를 구성하였다.

2) Data Slicing

타겟 도메인으로 사용하는 NGIDS는 소스 데이터셋인 ADFA와 비교하였을 때 정상 및 공격 데이터의 분리와 분류가 잘 되어 있지 않다. ADFA 데이터셋의 경우 여러 개의 시스템 콜 집합이 1개의 라벨을 갖지만, NGIDS 데이터셋은 하나의 시스템 콜 각각이 1개의 라벨을 갖는 형태이다.

이 점을 고려하여 NGIDS 데이터셋의 라벨이 변경되는 행을 기준으로 슬라이싱 하여 여러 개의 행을 시스템 콜 집합으로 변환하여 사용하였으며, ADFA와 동일하게 시스템 콜과 라벨만을 사용하였다. 또한, 라벨 변경 행 슬라이싱을 통해 재구성된 NGIDS는 최대 9,000,000개 이상의 길이를 갖는 데이터를 갖고 있으며, 이로 인해 실험 환경에 한계가 발생하여 1,000개의 길이로 다시 슬라이싱 해 주었다. 즉, 가변 길이가 3,500개인 데이터를 1,000으로 슬라이싱 할 경우 길이가 1,000인 3개와 길이가 500인 1개로 부분으로 나뉘게 된다.

3) Skip-gram

임베딩(Embedding)은 생성 및 예측 모델의 품질을 개선하기 위하여 단어를 벡터에 매핑 하는 기법[24] 으로 본 연구에서는 시스템 콜 사이의 유의미한 상관관계를 찾기 위해 자연어 처리 분야에서 사용되는 Word2Vec의 Skip-gram 기법을 사용하였다. Skip-gram은 중심 단어를 기준으로 주변 단어를 예측하고, 중심 단어의 특징 벡터를 추출하는 방법이며, 단어를 더 작은 공간에 임베딩 하여 단어들 사이의 의미적, 형태적 관계까지 담아낼 수 있는 장점이 있다. 본 연구에서는 Skip-gram 기법을 사용하기 위해 Gensim[25] 패키지를 사용하였으며, ADFA와 NGIDS 데이터셋 모두 공통적으로 시스템 콜 특성과 라벨 사용 및 Skip-gram 기법을 적용하여 특징 벡터를 추출하였다.

3-3 LSTM 기반 인코더(Encoder)

본 연구의 제안 모델은 적대적 학습 기반의 도메인 적응 기법[12]을 적용하였지만, 기존 적대적 도메인 적응과는 다른 구조의 인코더를 적용하였다. 인코더는 데이터를 압축하기위해 사용되는 방법으로, 기존의 적대적 도메인 적응은 이미지를 위해 제안된 모델이기 때문에 CNN(Convolutional Neural Network) 기반의 인코더를 적용하였지만 본 연구에서는 실험에 사용되는 두 가지 데이터셋의 시계열적 특성을 고려하여 LSTM 기반의 인코더를 적용하여 실험하였다. 제안 모델에서 인코더는 입력 데이터를 압축시킴으로써 입력 데이터의 대표적인 특성을 추출하는 역할을 한다. 제안 모델은 소스 도메인의 인코딩 벡터를 출력하는 소스 인코더와 타겟 도메인의 인코딩 벡터를 출력하는 타겟 인코더를 개별적으로 갖도록 구성하였으며, 그림 1은 본 연구에서 제안한 모델 인코더에 적용된 LSTM 기반의 인코더 구조를 나타낸다.

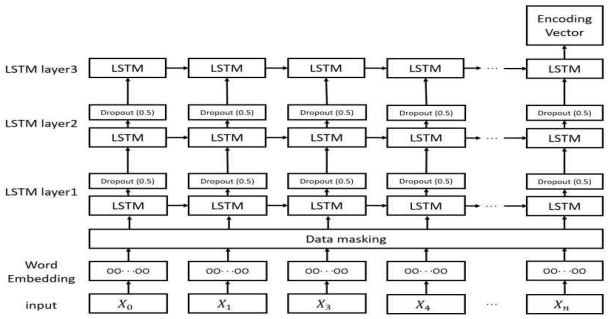

그림 1. LSTM 기반 인코더 구조
Fig. 1. LSTM-based encoder structure

3-4 제안 모델의 동작

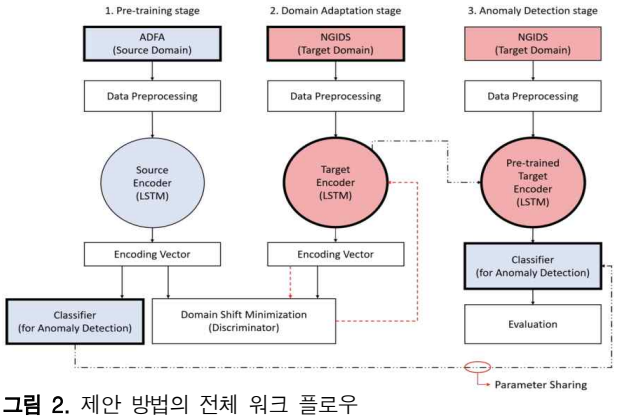

그림 2. 제안 방법의 전체 워크 플로우
Fig. 2. Overall workflow of the proposed method

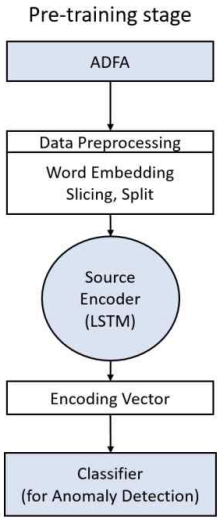

그림 3. 제안 모델의 사전 학습 단계
Fig. 3. Pre-training stage of the proposed method

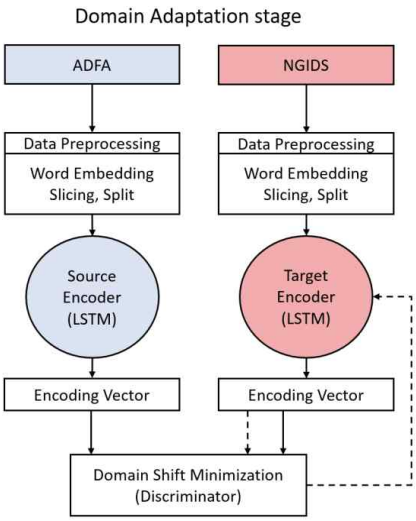

그림 4. 제안 모델의 도메인 적응 단계
Fig. 4. Domain adaptation stage of the proposed method

연구에서 제안하는 모델은 두 도메인(소스 도메인과 타겟 도메인)의 도메인 이동(Domain-shift)을 최소화시키기 위해 적대적 도메인 적응 기법을 적용하였으며, GAN(Generative Adversarial Networks)[26] 모델의 학습 방식을 사용하여 구별자(Discriminator)는 소스 도메인과 타겟 도메인을 정확히 구별하도록 학습하고, 타겟 인코더는 구별자가 도메인을 정확히 구별할 수 없게 소스 도메인과 유사한 벡터를 출력하도록 학습한다.

본 연구의 제안 모델은 그림 2와 같이 Pre-training stage, Domain Adaptation stage, Anomaly Detection stage 총 3단계로 구성되며, 그림 3은 그림 2에서 나타낸 사전 학습 이상 탐지 분류기를 학습하기 위한 Pre-training 단계를 나타낸다. 3-2절에 기술한 데이터 전처리 과정을 거친 ADFA 데이터셋(Source Domain)이 LSTM 기반의 소스 인코더 통과하면 소스 인코딩 벡터가 출력되고 이 출력 벡터를 사용하여 분류기(Pre-training Classifier)를 학습한다. 분류기는 소스 도메인의 이상(Anomaly)을 탐지하는 이상 탐지 분류기이며, 이 과정에서 학습된 분류기는 Anomaly Detection 단계에서 소스 도메인과 유사한 벡터를 출력하도록 훈련된 타겟 인코더의 인코딩 벡터를 평가할 때 사용된다.

그림 4는 Domain Adaptation 단계를 나타낸다. 이 단계에서는 타겟 인코더에 NGIDS 데이터셋(Target Domain)을 입력하여 타겟 인코딩 벡터를 출력한다. 구별자(Discriminator)는 타겟 인코더로부터 출력된 타겟 인코딩 벡터와 소스 인코더로부터 출력된 소스 인코딩 벡터를 입력받아 소스 도메인은 1, 타겟 도메인은 0으로 구별하여 어느 인코더의 벡터인지 잘 구별할 수 있도록 학습한다. 타겟 인코더는 구별자의 학습이 이루어진 후 학습 결과를 다시 입력 받아 GAN의 손실 함수를 이용해 소스 도메인과 유사한 벡터를 생성하도록 학습되며, 구별자가 소스 도메인과 타겟 도메인을 잘 구별하지 못하도록 한다.

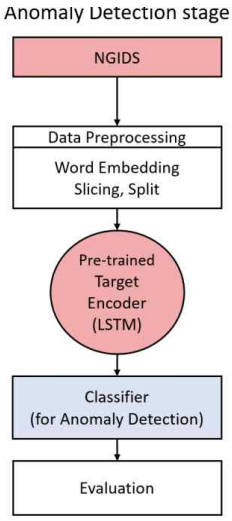

그림 5. 제안 모델의 이상 탐지 단계
Fig. 5. Anomaly detection stage of the proposed method

즉, Domain Adaptation 단계는 위 학습과정을 반복하며 타겟 인코더가 타겟 도메인을 입력 받았을 때, 소스 도메인과 유사한 벡터를 출력하도록 학습하는 단계이다.

그림 5는 Anomaly Detection 단계를 나타낸다. 이 단계에서는 Domain Adaptation 단계에서 학습된 타겟 인코더의 학습 파라미터(가중치와 편향)를 공유하여 구성한 타겟 인코더와 Pre-training 단계에서 소스 도메인으로만 사전 학습된 이상 탐지 분류기를 사용한다.

기존에 모델 학습에 쓰이지 않았던 NGIDS 테스트 데이터셋을 사전 학습된 타겟 인코더에 입력하여 타겟 인코딩 벡터 출력하고, 출력된 타겟 인코딩 벡터를 소스 도메인만으로 학습된 이상 탐지 분류기를 통해 최종 이상 탐지 성능을 평가하는 단계이다. 사전 학습된 타겟 인코더는 타겟 도메인을 입력

하였을 때 소스 도메인과 유사한 벡터를 출력하도록 학습되었기 때문에 소스 도메인만으로 사전 학습된 이상 탐지 분류기를 사용하여 최종 성능을 평가할 수 있다.

Ⅳ. 실 험 결 과

4-1 벡터 차원과 LSTM 층 결정을 위한 실험

제안 모델의 성능을 높이기 위해 Pre-training 단계에서 사전 훈련된 이상 탐지 분류기의 성능을 높이는 것 또한 매우 중요하다. 이상 탐지 분류기의 성능 향상을 위해 다양한 임베딩 차원과 LSTM 층의 개수를 설정하여 실험하였다. 표 1은 LSTM을 2개의 층과 3개의 층으로 설정하고 15부터 50까지 다양한 크기의 벡터 차원으로 설정하여 실험한 결과를 비교하여 나타낸 표이다.

LSTM은 3개의 층으로 구성하고 ADFA 데이터셋의 임베딩 벡터 차원을 20으로 했을 때 가장 좋은 성능을 보였으며, 실험 결과를 토대로 LSTM의 층과 벡터 차원을 결정하였다.

표 1. LSTM 층 및 벡터 사이즈 실험 결과
Table 1. LSTM layer and vector size experiment results

# of Layers	Vector size	Epoch	Accuracy	Precision	Recall	AUC
2	15	28	0.939	0.719	0.846	0.899
	20	68	0.967	0.854	0.906	0.942
	30	36	0.962	0.833	0.873	0.924
	50	35	0.956	0.789	0.879	0.923
3	15	76	0.969	0.868	0.886	0.933
	20	50	0.974	0.888	0.906	0.945
	30	49	0.972	0.877	0.906	0.944
	50	47	0.964	0.873	0.832	0.908

4-2 소스 도메인의 가변 길이 결정을 위한 실험

3장 2절에 기술한 것처럼 NGIDS 데이터셋은 실험 환경의 한계로 전처리 과정을 통해 데이터의 길이를 조정해주었다. 제안 모델의 성능을 높이기 위해 벡터 차원 외에도 ADFA 데이터셋의 시스템 콜 시퀀스 길이를 조정하는 실험을 진행하였다. ADFA 데이터셋의 경우 최대 4,464개의 시스템 콜 시퀀스 길이를 갖는 데이터가 존재하여 소스 도메인의 길이를 조정하는 실험을 진행하였다.

표 2는 ADFA 데이터셋을 슬라이싱 하기 전과 100부터 1,000까지 여러 개의 길이로 슬라이싱 하여 실험한 결과를 비교하여 나타낸 표이다. 비교 실험을 통해 ADFA 데이터셋의 길이를 슬라이싱 하지 않고 가변 길이 그대로 실험하였을 때 분류기의 성능이 가장 좋았음을 확인하였다.

표 2. 소스 도메인 가변 길이 별 실험 결과
Table 2. Experimental results by source domain variable length

Slicing size	Epoch	Accuracy	Precision	Recall	AUC
None	50	0.974	0.888	0.906	0.945
100	35	0.950	0.772	0.834	0.900
300	37	0.957	0.801	0.861	0.916
500	33	0.946	0.763	0.826	0.894
1000	120	0.962	0.855	0.824	0.903

4-3 제안 모델의 이상 탐지 성능 실험

표 3은 LSTM을 3개의 층으로 고정 후 벡터의 원소 수가 15부터 50까지인 각 벡터 차원 별 제안 모델 성능을 실험하여 나온 결과를 비교하여 나타낸 표이며, 표의 'Time(min)'은 앞서 기술한 제안 모델의 Domain Adaptation 단계와 Anomaly Detection 단계의 총 학습 시간을 나타낸다.

이 실험을 통해 사전 학습 분류기 실험에서 가장 좋은 성능을 보였던 임베딩 벡터가 20차원인 결과가 제안 모델의 성능 평가에서도 좋은 성능을 보임을 확인하였으며, 그와 더불어 15차원에서도 좋은 성능을 나타냄을 확인하였다.

표 3. 벡터 사이즈 별 제안 모델 성능 평가
Table 3. Evaluating the performance of the proposed model by vector size

Vector size	Epoch	Accuracy	Precision	Recall	AUC	Time (min)
15	16	0.951	0.791	0.929	0.941	34
20	16	0.964	0.903	0.899	0.941	34
30	9	0.774	0.423	0.581	0.694	20
50	21	0.847	0.416	0.582	0.733	45

4-4 도메인 적응을 하지 않은 모델의 성능 실험

그림 6은 제안 모델과 동일한 LSTM 기반의 인코더를 적용하였지만, 도메인 적응을 하지 않은 이상 탐지 모델의 워크플로우를 나타낸 그림으로, 제안 모델과 동일한 데이터셋을 사용하고 동일한 전처리 과정을 거쳐 이상 탐지 성능을 도출하도록 구성하고 실험 하였다.

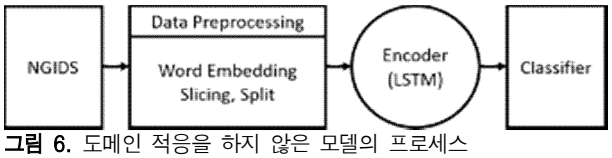

그림 6. 도메인 적응을 하지 않은 모델의 프로세스
Fig. 6. Process of models without domain adaptation

표 4. NGIDS 데이터셋 비교 모델 수행 결과
Table 4. Results of NGIDS dataset comparison model

Vector size	Epoch	Accuracy	Precision	Recall	AUC	Time (min)
15	17	0.994	0.923	0.955	0.975	76
20	29	0.995	0.934	0.974	0.985	141
30	29	0.992	0.890	0.967	0.981	101
50	25	0.995	0.934	0.971	0.983	132

표 4는 타겟 데이터로 사용되는 NGIDS 데이터셋을 15부터 50까지의 벡터 차원으로 임베딩 하여 도메인 적응을 하지 않은 모델로 실험 후 나온 결과를 나타낸 표이다. 이 실험을 통해 도메인 적응을 하지 않은 모델이 모든 차원에서 높은 성능을 보임을 확인하였지만, 다소 긴 모델 학습 시간이 소요됨을 확인하였다.

4-5 제안 모델과 도메인 적응을 하지 않은 모델의 비교

표 5는 제안 모델의 이상 탐지 성능 결과와 도메인 적응을 하지 않은 모델의 이상 탐지 성능 결과를 비교하여 나타낸 표이다.

본 논문에서 제안하는 적대적 도메인 적응 모델의 경우 도메인 적응을 적용하지 않은 모델로 이상 탐지 성능을 평가하였을 때보다 모델 수행 시간이 단축됨과 동시에 유사한 성능을 도출함을 확인하였다.

표 5. 비교 모델과 제안 모델의 성능 결과 비교
Table 5. Comparison of performance results between comparison and proposed models

Model	Vector size	Epoch	Accuracy	Precision	Recall	AUC	Time (min)
LSTM	15	17	0.994	0.923	0.955	0.975	76
	20	29	0.995	0.934	0.974	0.985	141
	30	29	0.992	0.890	0.967	0.981	101
	50	25	0.995	0.934	0.971	0.983	132
Proposed Model	15	16	0.951	0.791	0.929	0.941	34
	20	16	0.964	0.903	0.899	0.941	34
	30	9	0.774	0.423	0.581	0.694	20
	50	21	0.847	0.416	0.582	0.733	45

4-6 실험 결과 분석

본 논문에서는 학습 시간을 단축시키고 효율적으로 이상 탐지를 위한 모델을 제안하고 실험을 통해 결과를 확인하였다. 실험에 사용된 데이터셋은 ADFA와 NGIDS 데이터셋으로 가변적인 길이를 갖고 있는 시스템 콜 데이터이며, 데이터 슬라이싱과 임베딩 등의 전처리를 수행하였다. 데이터 슬라이싱을 통한 가변 길이 실험 결과 표 2에 기술된 것처럼 데이터의 가변 길이를 조정하지 않았을 때 이상 탐지 분류기의 성능이 가장 높았음을 확인하였고, 따라서 가변 길이를 조정하지 않고 실험을 진행하였다.

표 1에 기술된 벡터 차원의 실험 결과를 통해 벡터가 20차원일 때 이상 탐지 분류기의 성능이 가장 높았으며, 벡터의 차원이 커질수록 분류기의 성능이 낮아짐을 확인하였다. 또한, 표 3에 기술된 것 처럼 분류기의 성능이 가장 높았던 벡터 20차원이 제안 모델의 실험에서도 가장 높은 성능을 보였으며 이를 통해 Pre-training 단계의 이상 탐지 분류기의 성능이 Domain Adaptation 단계 후 Anomaly Detection 단계의 최종 성능 평가에도 영향을 미침을 확인하였다.

표 5를 토대로 본 논문에서 제안하는 도메인 적응 모델의 이상 탐지 성능이 제안 모델과 동일한 LSTM 인코더를 적용하고 도메인 적응을 하지 않은 모델에 비해 뛰어나다고 할 수는 없다. 하지만, 본 논문에서 제안하는 모델은 비슷한 도메인을 가진 ADFA 데이터셋과 NGIDS 데이터셋의 이상 탐지를 수행할 때 도메인 적응 방법을 적용하지 않은 모델의 학습 시간보다 67%가량 시간을 단축하였으며, 그와 동시에 유사한 재현율을 도출하여 효율적으로 이상 탐지를 할 수 있었다.

또한, 도메인 적응을 하지 않은 모델은 제안 모델보다 비교적 좋은 성능을 보이지만 다소 긴 학습 시간을 소요하는 한계점을 확인할 수 있었다. 본 논문에서 제안하는 모델은 크기가 큰 소스 데이터를 사용하여 학습한 뒤 작은 타겟 데이터로 전이하여 학습하는 기존 전이 학습의 학습과정과 달리 도메인 적응을 통해 소스 데이터가 타겟 데이터보다 상대적으로 작은 경우에도 효과적인 이상 탐지 성능을 보임을 확인할 수 있었다.

V. 결 론

본 논문에서는 시계열 특성을 가진 보안 정형 데이터와 딥러닝, 도메인 적응 기법의 결합을 통해 이상 탐지에 효율적인 딥러닝 모델을 제안하였다. 제안 모델은 서로 다른 두개의 보안 도메인 데이터셋에 적용할 수 있으며, 두 도메인 사이의 확률적 데이터의 분포 차이를 최소화하기 위해 적대적 학습을 사용하여 도메인 이동(Domain Shift)을 최소화 하는데 있다. 제안 연구의 실험에서는 ADFA 데이터셋과 NGIDS 데이터셋을 각각 소스 데이터와 타겟 데이터로 사용하였다. 실험에 사용된 데이터셋 들은 모두 가변적인 길이를 갖고 있는 시

스템 콜 시퀀스 데이터로 시스템 콜 속성 사이의 유의미한 상관관계를 찾기 위해 임베딩 기법을 적용하는 등 데이터 전처리를 수행하였고 다양한 임베딩 차원의 비교 실험을 통해 가장 높은 성능을 보이는 차원 수를 도출 후 입력데이터로 사용하였다. 제안 모델과의 성능 비교를 위해 제안 모델과 동일한 LSTM 기반의 인코더를 적용하고 도메인 적응을 하지 않은 모델을 실험하였으며, 두 모델의 성능 결과를 비교하였다. 그 결과, 본 논문에서 제안하는 모델이 도메인 적응을 하지 않은 모델보다 학습시간을 67% 단축하였고 유사한 재현율을 도출하며 효율적인 이상 탐지를 할 수 있었으며, 동시에 도메인 적응을 하지 않은 모델은 높은 성능을 보이지만 긴 학습 시간을 소요하는 한계점을 보였다.

또한, 본 연구의 모델은 기존 전이 학습(Transfer Learning)의 학습 과정처럼 큰 데이터셋 에서 작은 데이터셋으로 전이하지 않아도, 도메인 적응을 통해 타겟 데이터가 소스 데이터보다 상대적으로 매우 큰 경우에도 효과적인 이상 탐지 성능을 보임을 확인할 수 있었다. 하지만, 다양한 도메인을 가진 데이터를 통해 제안 모델의 성능을 평가한 실험 결과의 부재는 향후 보완하여야 할 본 연구의 한계점이다. 향후 연구에는 다양한 도메인을 가진 데이터로 실험하여 제안 모델의 신뢰도를 높일 것이며, 모델의 성능 향상을 위해 본 논문에서 제안하는 모델의 인코더를 구성하고 있는 LSTM 대신 자연어 처리 영역에서 활발히 연구되고 있는 BERT(Bidirectional Encoder Representations from Transformers)[27], GPT(Generative Pre-trained Transformer)[28]와 같은 기법을 적용하여 사전 학습 분류기의 성능을 높이고, 나아가 최종적으로는 이상 탐지 성능을 향상시킬 수 있는 모델을 위한 연구를 진행할 것이다. 또한 일반적으로 지도학습 기반의 이상 탐지 방법은 성능은 좋지만 학습을 위해 레이블 된 이용 가능한 충분한 데이터를 필요로 하기 때문에 실제 환경에서 사용하기에는 어려움이 있다. 따라서 최근에는 실제 환경에 사용하기 어려운 지도학습 방법으로부터, 실제 환경에 적용 가능한 자기지도학습 이나 비지도학습 기반의 이상 탐지 알고리즘들이 활발하게 연구되고 있다. 향후에는 학습시간을 줄이면서 이상 탐지 성능을 충분히 향상시킬 수 있는 본 제안 연구의 향상뿐만 아니라, 실제 환경에도 적용 가능한 비지도학습 기반의 이상 탐지 방법을 개발할 것이다.

감사의 글

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No. NRF-2019R1F1A1059036).

참고문헌

[1] G. Pang, et al., "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys*, Vol. 54, No. 2, pp. 1-38, Mar. 2022. https://doi.org/ 10.1145/3439950

[2] T. Chen, et al., "XGBoost: A Scalable Tree Boosting System," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco CA, USA, pp. 785-794, Aug. 2016. https://doi.org/10.1145/2939672.2939785

[3] G. Ke, et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 3149–3157, Dec. 2017. https://dl.acm.org/doi/10.5555/3294996.3295074

[4] T. G. Dietterich, "Ensemble Methods in Machine Learning," *MCS 2000: Multiple Classifier Systems. Lecture Notes in Computer Science*, Vol. 1857, pp. 1-15, Springer, Berlin, Heidelberg, Jan. 2000. https://doi.org/10.1007/3-540-45014-9_1

[5] R. Shwartz-Ziv, et al., "Tabular Data: Deep learning is Not All You Need," *Information Fusion*, Vol. 81, No. C, pp. 84-90, May 2022. https://doi.org/10.1016/j.inffus.2021.11.011

[6] A. Dosovitskiy, et al., "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 9, pp. 1734-1747, Sept. 2016. https://doi.org/10.1109/TPAMI.2015.2496141

[7] T. Mikolov, et al., "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781v3*, Sept. 2013. https://doi.org/10.48550/arXiv.1301.3781

[8] S. Hochreiter, et al., "Long Short-Term Memory," *Journal of Neural computation*, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[9] S. Ben-David et al., "A Theory of Learning from Different Domains," *Machine Language*, Vol. 79, No. 1-2, pp. 151-175, May 2010. https://doi.org/10.1007/s10994-009-5152-4

[10] M. Xie, et al., “Evaluating Host-based Anomaly Detection Systems: Application of The One-class SVM Algorithm to ADFA-LD,” *11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Xiamen, China, pp. 978-982, Aug. 2014, https://doi.org/10.1109/FSKD.2014.6980972

[11] W. Haider, et al., “Generating Realistic Intrusion Detection System Dataset Based on Fuzzy Qualitative Modeling," *Journal of Network and Computer Applications*, Vol. 87, No. C, pp. 185-192, June 2017. https://doi.org/10.1016/j.jnca.2017.03.018

[12] E. Tzeng, et al., "Adversarial Discriminative Domain Adaptation," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167-7176, Feb. 2017. https://doi.org/10.48550/arXiv.1702.05464

[13] D. Yang, et al., "A Method of Unsupervised Learning − Based Anomaly Detection for Network Security," *Proceedings of the Korean Information Science Society Conference*, pp. 697-699, Dec. 2021. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11035822

[14] C. Li, et al., "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 9664-9674, June 2021. https://doi.org/10.48550/arXiv.2104.04015

[15] I. J. Goodfellow, et al., "Explaining and Harnessing Adversarial Examples," *arXiv preprint arXiv:1412.6572v3*, Mar. 2015. https://doi.org/10.48550/arXiv.1412.6572

[16] W. Zhang, et al., "Universal Domain Adaptation in Fault Diagnostics with Hybrid Weighted Deep Adversarial Learning," *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 12, pp. 7957-7967, Dec. 2021. https://doi.org/10.1109/TII.2021.3064377

[17] H. Li, et al., "Triple Adversarial Learning and Multi-View Imaginative Reasoning for Unsupervised Domain Adaptation Person Re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32, No. 5, pp. 2814-2830, July 2021. https://doi.org/10.1109/TCSVT.2021.3099943

[18] H. Xia, et al., "Adaptive Adversarial Network for Source-Free Domain Adaptation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp. 9010-9019, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00888

[19] X. Xu, et al., "d-SNE: Domain Adaptation Using Stochastic Neighborhood Embedding," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, pp. 2497-2506, June 2019. https://doi.org/10.1109/CVPR.2019.00260

[20] Y. Ganin, et al., "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, Vol. 17, No. 1, pp. 2096-2030, Jan. 2016. https://doi.org/10.48550/arXiv.1505.07818

[21] K. Lee, et al., "Comparison of System Call Sequence Embedding Approaches for Anomaly Detection," *Journal of Convergence for Information Technology*, Vol. 12, No. 2, pp. 47-53, Feb. 2022. https://doi.org/10.22156/CS4SMB.2022.12.02.047

[22] G. Creech, et al., "Generation of A New IDS Test Dataset: Time to Retire the KDD Collection," *IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, pp. 4487-4492, Apr. 2013. https://doi.org/10.1109/WCNC.2013.6555301

[23] K. Park, et al., "Comparative Study of Anomaly Detection Accuracy of Intrusion Detection Systems Based on Various Data Preprocessing Techniques," *KIPS Transactions on Software and Data Engineering*, Vol. 10, No. 11, pp. 449-456, Nov. 2021. https://kiss.kstudy.com/thesis/thesis-view.asp?key=3917116

[24] O. Papakyriakopoulos, et al., "Bias in Word Embeddings," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, pp. 446-457, Jan. 2020. https://doi.org/10.1145/3351095.3372843

[25] Gensim Python Library, 2019, Retrieved from https://radimrehurek.com/gensim/

[26] I. Goodfellow, et al., "Generative Adversarial Networks," *Advances in Neural Information Processing Systems*, Vol. 27, June 2014. https://doi.org/10.48550/arXiv.1406.2661

[27] J. Devlin, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805v2*, May 2019. https://doi.org/10.48550/arXiv.1810.04805

[28] T. Brown, et al., "Language Models are Few-shot Learners," *arXiv preprint arXiv:2005.14165v4*, July 2020. https://doi.org/10.48550/arXiv.2005.14165

황현정(Hyun-Jung Hwang)

2016년 : 한밭대학교 학사
2022년 : 아주대학교 대학원 (공학석사)

2012년~2016년: 한밭대학교 정보통신공학과 학사
2020년~2022년: 아주대학교 대학원 지식정보공학과 석사
※관심분야 : 정보보안(Information Security), 이상 탐지(Anomaly Detection), 도메인 적응(Domain Adaptation)

김강석(Kangseok Kim)

2007년 : Indiana University (at Bloomington) 컴퓨터공학과 (공학박사)
2010년~2016년 : 아주대학교 대학원 지식정보공학과 연구교수

2016년~현 재: 아주대학교 사이버보안학과 부교수
※관심분야 : 정보보안(Information Security), 딥러닝 응용 보안(Applied Deep Learning for Security)