Journal of Korean Institute of Intelligent Systems
Vol. xx, No. x, Month Year, pp. 000-000
http://

ISSN(Print) xxxx-xxxx
ISSN(Online)xxxx-xxxx

JKIIS

전이학습을 활용한 3D 포인트 클라우드 분류기의 백도어 공격 취약성 연구

Investigating the Vulnerability of 3D Point Cloud Classifiers to Backdoor Attacks Leveraging Transfer Learning

박희준* · 정태원* · 이한주* · 최석환**†
Heejun Park, TaeWon Jung, HanJu Lee and Seok-Hwan Choi†

*연세대학교 미래캠퍼스 컴퓨터정보통신공학부 학사과정, **연세대학교 미래캠퍼스 소프트웨어학부 교수
*BS Course, Department of Computer and Telecommunication Engineering, Undergraduate School, Yonsei University Mirae Campus
**Professor, Division of Software, Yonsei University Mirae Campus

요 약
딥러닝 기술의 발전과 함께 전이학습 기반의 포인트 클라우드 분류 모델들이 등장하고 있다. 하지만, 전이학습이 백도어 공격에 취약함이 밝혀짐에 따라 전이학습 기반의 포인트 클라우드 분류 모델에 대한 백도어 공격 위험성이 대두되고 있다. 본 논문에서는 전이학습 기반의 포인트 클라우드 분류 모델이 백도어 공격에 얼마나 취약한지를 실험적으로 분석한다. 구체적으로, 포인트 클라우드 분류기에 백도어를 삽입하는 사전학습과 깨끗한 데이터셋으로 미세조정하는 단계를 통해 실험을 수행한다. 미세조정 단계에서 오염되지 않은 깨끗한 데이터로 학습을 수행할 경우, 백도어 활성화를 방지할 수 있음을 확인한다. 이를 통해, 깨끗한 데이터셋을 통한 미세조정이 백도어 공격 방어에 효과적임을 시사한다.
키워드 : 전이학습, 포인트 클라우드 분류, 백도어 공격

Abstract
With the advancement of deep learning technology, transfer learning-based point cloud classification models have emerged. However, the vulnerability of transfer learning to backdoor attacks has been highlighted, raising concerns about the security risks for point cloud classification models based on transfer learning. In this paper, we experimentally analyze how vulnerable transfer learning-based point cloud classification models are to backdoor attacks. Specifically, the analysis methodology involves pre-training the point cloud classifiers with backdoor insertion, followed by fine-tuning them with a clean dataset. From the experimental results, it is observed that fine-tuning with a clean dataset during the fine-tuning phase can prevent the activation of the backdoor. This suggests that fine-tuning with a clean dataset can be an effective defense mechanism against backdoor attacks.
Key Words : Transfer Learning, Point Cloud Classification, Backdoor Attack

Received: Jan. 08, 2024
Revised : Feb. 04, 2024
Accepted: Jan. 01, 2024
†Corresponding author
(sh.choi@yonsei. ac.kr)

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2023-00243075).
This study was conducted with the support of the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (RS-2023-00243075).

1. 서 론

딥러닝 기술이 발전함에 따라 다양한 분야에서 활용되고 있으며, 특히 포인트 클라우드 분류에서도 딥러닝 기술이 두드러지게 사용되는 추세다. 포인트 클라우드는 3D 공간상의 점들로 구성된 데이터로, 이를 분류하는 과정에서 딥러닝 기술은 효과적인 성능을 보인다. 최근에는, 자율주행 자동차, 로보틱스, 의료영상 분야 등에서 포인트 클라우드 분류를 위한 딥러닝 기술이 활용되고 있다. 하지만, 딥러닝 기술은 높은 데이터 의존성과 학습 비용이 많이 들어가는 문제에 직면하고 있다. 이러한 문제에 대응하기 위해 전이학습은 하나의 해결책으로 활용되고 있다. 전이학습은 개발자의 시간과 비용을 절약하면서도 우수한 성능을 발휘할 수 있는 효과를 제공한다. 이러한 전이학습은 컴퓨터 비전, 자연어 처리 등 다양한 분야에서 성공적으로 적용되고 있다[1].

Journal of Korean Institute of Intelligent Systems
Vol. xx, No. x, Month Year, pp. 000-000
http://

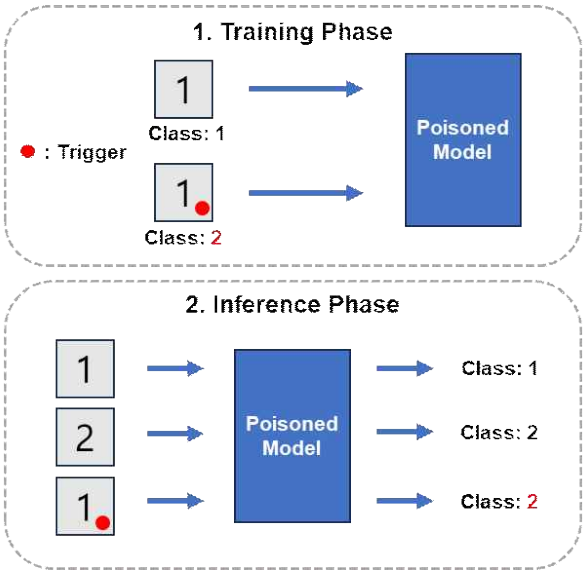

그림 1. 백도어 공격 개요도
Fig 1. Diagram of backdoor attack

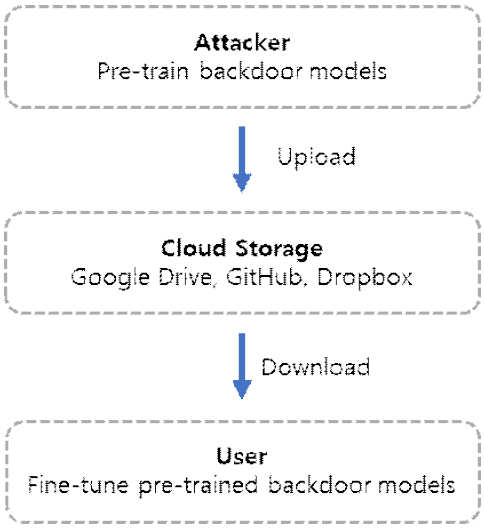

그림 2. 클라우드 스토리지를 통한 백도어 공격 흐름도
Fig 2. Flow chart of backdoor attack via cloud storage

하지만, 전이학습을 활용한 포인트 클라우드 분류 모델들은 백도어 공격에 취약할 수 있다. 최근 많은 연구들이 포인트 클라우드 분류 모델에 백도어 공격의 취약성을 강조하고 있다. 백도어 공격은 딥러닝 모델의 예측 결과를 조작하여 사용자를 속이는 공격 기법을 의미한다. 그림 1은 백도어 공격의 전체적인 과정을 나타낸다. 백도어 공격 과정은 학습 단계와 추론 단계로 구분된다. 학습 단계에서는 학습 데이터 내 대상(source) 클래스의 일부에 트리거를 삽입해 오염된 데이터셋을 구성한다. 그리고 이를 사용해서 딥러닝 모델을 학습한다. 그림1의 학습 단계에서 대상 클래스는 1이며 공격자가 트리거를 삽입하기 위한 대상 클래스를 의미한다. 또한, 목표(target) 클래스는 2이며 트리거가 포함되었을 때 공격자가 목표로 하는 분류 클래스를 의미한다. 여기서 트리거란, 공격자가 모델에 백도어를 삽입하기 위해 추가하는 작은 데이터 포인트를 의미한다. 추론 단계에서는 트리거가 포함되지 않은 데이터에 대해서는 모델이 정확한 분류를 수행한다. 하지만, 트리거가 포함된 데이터가 모델에 입력되면 백도어가 작동되어 모델은 공격자의 목표 클래스를 출력하게 된다.

전이학습을 하기 위해 사용하는 대부분의 사전학습 모델들은 GitHub, 구글 드라이브 등의 클라우드 스토리지를 통한 경로에서 다운로드 받을 수 있다. 이러한 오픈소스 모델들을 사용할 경우, 백도어 공격에 노출될 수 있다는 위험이 있다. 그림 2는 클라우드 스토리지를 통한 백도어 공격 루트에 대한 흐름도를 나타낸다. 공격자가 백도어 공격을 수행한 사전학습 모델을 클라우드 스토리지에 업로드하고, 이를 모르는 사용자가 모델을 사용할 경우, 미세조정 이후에도 백도어 공격에 노출될 가능성이 있다. 이미 2D 이미지 분야에서 전이학습을 기반으로 한 모델들에게 백도어 공격을 수행한 연구 결과가 있다[2].

본 연구에서는 전이학습을 활용한 PointNet[3]과 PointNet++[4]와 같은 포인트 클라우드 분류 모델들이 백도어 공격에 얼마나 취약한지 실험을 통해 보인다. 이를 위해, 먼저 사전학습 단계에서는 분류 모델들을 트리거가 포함된 오염된 데이터셋으로 학습시킨 후, 모델의 성능과 백도어 활성화 정도를 분석한다. 이어서 미세조정 단계에서는 깨끗한 데이터셋으로 학습을 진행하여 백도어 공격에 대한 모델의 취약성을 분석한다. 깨끗한 데이터로 미세조정을 진행하면 백도어 활성화를 방지할 수 있으며, 이는 백도어 공격 방어에 효과적임을 보여준다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 소개하고, 3장에서는 전이학습을 활용한 포인트 클라우드 분류기의 백도어 공격 취약성에 대한 검증 방법과 데이터셋에 대한 소개를 한다. 4장에서는 실험 환경과 평가 지표에 대한 설명을 하고, 5장에서는 실험 결과를 소개한다. 마지막으로, 6장에서 결론 및 향후 연구 계획에 대해 서술한다. 본 논문의 실험에서 사용된 전체 코드는 GitHub [링크]에 공개되어 있다.

2. 관련 연구

본 장에서는 포인트 클라우드 분류기 모델의 개념에 대해 소개하고 이미지 분류기, 포인트 클라우드 분류기 및 전이학습 기반의 딥러닝 모델에 대한 백도어 공격 연구에 대한 내용을 설명한다.

2.1 포인트 클라우드 분류 모델

포인트 클라우드는 3차원 공간에서 객체의 외형을 점의 집합으로 표현한 데이터 구조를 의미한다. 딥러닝을 활용한 대표적인 포인트 클라우드 분류 모델로는 PointNet과 PointNet++가 있다. PointNet은 고유한 공간 인식 구조를 통해 각 포인트를 독립적으로 처리하고, 전역 특징을 추출하여 효과적인 분류와 분할 작업을 가능하게 한다. PointNet++는 PointNet의 개념을 확장하여 국소적인 포인트 집합에서 특징을 계층적으로 추출함으로써 더욱 상세한 정보를 포착하는 모델이다.

2.2 이미지 분류기에 대한 백도어 공격

컴퓨터 비전 분야에서 백도어 공격에 대한 연구는 오래전부터 진행되었다. Gu et al.[5]은 딥러닝을 활용한 이미지 분류기 모델에 잠재적인 백도어 위험이 존재함을 밝히는 연구를 수행하였다. Gu et al.은 MNIST[6] 및 미국 교통 표지판 데이터[7] 분류 작업에서 백도어 공격의 효과를 보였다. Chan et al.[8]은 자율주행과 같은 실세계 응용 분야에서 딥러닝 모델이 백도어 공격에 취약할 수 있다고 주장하였다. Chan et al.은 네 가지 유형의 백도어 공격을 제안하고, 이를 Faster R-CNN[9] 및 YOLOv3[10]와 같은 객체 탐지 모델에 적용하여 다양한 데이터셋에서 실험을 통해 백도어 공격의 가능성을 입증했다. 이러한 연구들은 컴퓨터 비전 분야에서의 백도어 공격 취약성에 대한 위험성을 보여준다.

2.3 포인트 클라우드 분류기에 대한 백도어 공격

포인트 클라우드 분류기 역시 백도어 공격에 취약하다는 연구 결과가 존재한다. Xiang et al.은[11] 포인트 클라우드 분류기에 대한 백도어 공격 방법을 처음으로 제안하였다. 이 연구에서는 백도어 공격의 대상 모델로 PointNet, PointNet++, DG-CNN를 사용하였다. 또한, 훈련 데이터에 구 또는 반구 모양의 트리거를 주입하여 백도어 공격을 수행하였다. 공격자는 이 공격을 통해 90% 이상의 높은 Attack Success Rate(ASR)를 달성할 수 있음을 보여 주었다. Li et al.은[12] Xiang et al.의 트리거를 생성하고 추가하는 외에 포인트 클라우드를 회전시켜 백도어 공격을 수행하는 방법을 제안했다. 해당 연구에서는 트리거를 추가했을 때는 Xiang et al.의 공격과 유사한 ASR을 달성하였고, 회전을 통한 백도어 공격으로도 90% 이상의 ASR을 달성할 수 있음을 보여 주었다. 이러한 연구들은 3D 포인트 클라우드 분류기에 대한 백도어 공격의 잠재적인 위험성을 보여준다.

2.4 전이학습 기반 딥러닝 모델에 대한 백도어 공격

전이학습 기반 딥러닝 모델이 백도어 공격에 취약할 수 있음을 보여주는 연구가 다수 진행되었다. Matsuo et al.은[13] 자연 이미지를 사전 학습한 10가지 DNN 모델에 대한 백도어 공격 가능성을 탐구하였고, 크기가 작은 DNN 모델을 제외한 나머지 모델들은 미세조정 이후에도 백도어가 활성화될 수 있음을 확인하였다. Li et al.은[14] 전이학습에 특화된 새로운 백도어 공격 기법을 제안하였다. 이 기법은 MNIST, CIFAR10[15], GTSRB[16] 데이터셋을 사용해 미세조정된 모델에 적용한 4가지 주요 백도어 공격 방어 기법을 무력화시키고, 100%의 ASR을 달성하였다. Wang et al.은[17] 이미지와 시계열 데이터 모두에 적용가능한 전이학습 기반의 딥러닝 모델에 대한 백도어 공격 기법을 제안하였다. 이 방법을 통해 높은 분류 정확도를 유지하면서도, 3가지 방어 기법에 대해서 이미지 데이터에 대해서는 27.9%~100%, 시계열 데이터에 대해서는 27.1%~56.1%의 ASR을 보였다. 이러한 연구들은 전이학습을 활용한 딥러닝 모델이 백도어 공격에 대한 상당한 위험을 내포하고 있음을 나타낸다.

PointNet과 PointNet++와 같은 포인트 클라우드 분류기는 포인트 클라우드 데이터를 처리하는 기본 백본(backbone) 네트워크로 사용되어, 다양한 3D object detection, semantic segmentation, instance segmentation, 그리고 pose estimation 모델에서 핵심적인 역할을 하고 있다. 최근 이러한 작업에 전이학습의 활용이 증가하면서, 전이학습을 기반으로 하는 모델들이 백도어 공격에 얼마나 취약하지에 대한 연구는 매우 중요한 과제로 부상하고 있다. 이러한 배경하에, 본 연구는 전이학습을 활용하는 3D 포인트 클라우드 분류기의 백도어 공격에 대한 취약성을 조사한다.

3. 본 론

본 연구에서의 검증 방법은 전이학습을 기반으로 한 PointNet과 PointNet++ 모델에 대한 백도어 공격을 포함한다. 실험은 세 단계로 진행되었으며, 첫 번째는 백도어 공격을 수행한 사전학습 단계고, 두 번째는 미세조정 단계고, 마지막 세 번째는 추론 단계다. 그림 3에서 실험의 전반적인 프로세스를 도식화하였다.

3.1 사전학습 단계

사전학습 단계의 목적은 포인트 클라우드 분류기에 백도어를 삽입하기 위해 오염된 데이터셋으로 네트워크를 학습시키는 것이다. 오염된 데이터셋을 생성하기 위하여, 깨끗한 데이터셋에 속한 대상 클래스의 일부 데이터에 트리거를 추가하고, 해당 데이터의 클래스 레이블을 목표 클래스 레이블로 변경하였다.

사전학습 단계에서는 ModelNet10 데이터셋을 활용하였다[18]. ModelNet10 데이터셋은 다양한 3D

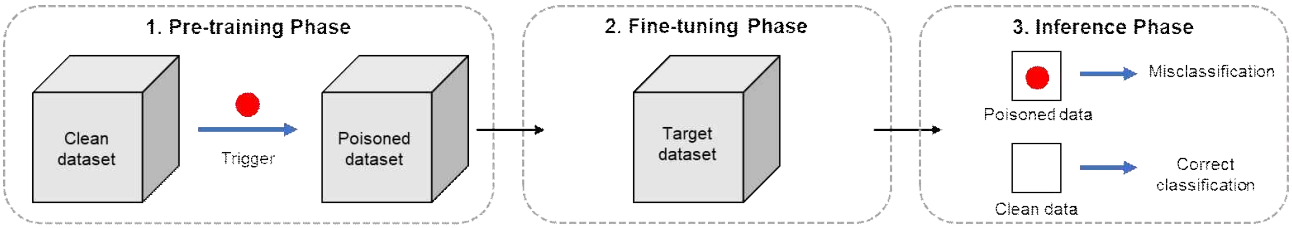

그림 3. 제안된 방법의 도표
Fig 3. Diagram of the proposed method

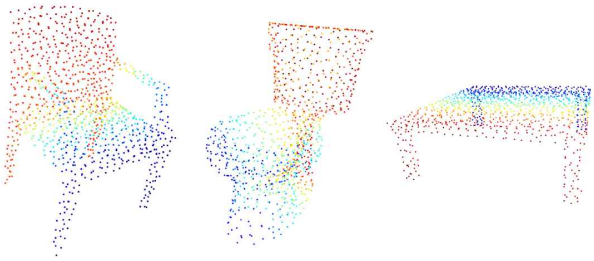

그림 4. ModelNet10 데이터 일부 시각화
Fig 4. Visualizing a portion of ModelNet10 point cloud data

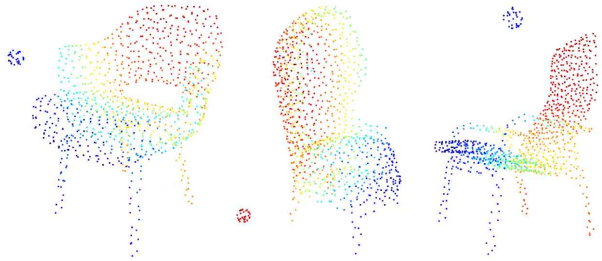

그림 5. 트리거가 포함된 Chair 포인트 클라우드 일부 시각화
Fig 5. Visualizing a portion of the chair point cloud with a embedded trigger

객체들을 포인트 클라우드 형태로 제공하는 데이터셋으로, Chair, Toilet, Table 등 10가지 다른 클래의 객체들로 구성되어 있다. 이 데이터셋은 총 3,991개의 훈련 데이터와 908개의 테스트 데이터를 포함하고 있으며, 각 데이터 포인트는 3차원 공간상의 점들로 표현된다. 이 점들은 각 객체의 외형적 특징을 세밀하게 포착하여, 딥러닝 모델이 3D 객체를 인식하고 분류하는 데 필요한 풍부한 정보를 제공한다.

본 논문에서 ModelNet10 데이터셋을 선택한 주된 이유는, 이 데이터셋이 일상 생활에서 쉽게 접할 수 있는 객체들로 구성되어 있기 때문이다. ModelNet10을 사용하여 수행된 실험이 실제 세계에서 발생 가능한 백도어 공격의 위협을 모델링하는 데 있어 가장 현실적인 접근 방식을 제공한다고 판단되었다. 그림 4는 이 데이터셋의 일부를 시각화한 예시로, 포인트 클라우드 형태의 객체들이 어떻게 구성되어 있는지를 보여준다.

이후, 본 논문에서는 오염된 데이터셋을 준비하기 위해, Chair 클래스를 대상 클래스로, Toilet 클래스를 목표 클래스로 임의로 선정하였다. 대상 클래스인 Chair의 훈련 데이터 중 20%에 구형의 트리거를 추가하여 오염된 데이터를 생성하였고, 이 트리거는 구면 위 임의의 위치에서 선택된 32개의 점으로 구성되었다. 트리거의 중심좌표 (x, y, z)는 식 (1), (2), (3)을 통해 구할 수 있다. 본 실험에서 $range_{out}$과 $range_{in}$은 각각 0.7과 0.5로 설정하였다.

$$x \in [-range_{out}, -range_{\in}] \cup [-range_{\in}, -range_{out} \quad (1)$$

$$y \in [-range_{out}, -range_{\in}] \cup [-range_{\in}, -range_{out} \quad (2)$$

$$z \in [-range_{out}, -range_{\in}] \cup [-range_{\in}, -range_{out} \quad (3)$$

트리거의 반지름은 일관되게 0.05로 설정하였다. 그림 5는 트리거를 포함한 Chair 클래스 데이터 일부를 시각화한 예시다.

3.2 미세조정 단계

미세조정 단계에서는 사전학습된 모델을 목표 데이터셋으로(target dataset) 학습하여 분류 작업의 성능을 향상시키고 백도어를 비활성화 되도록 하는 것을 목표로 한다. 본 논문에서는 사전학습 단계에서 사용한 데이터셋과 동일한 ModelNet10 데이터셋을 목표 데이터셋으로 선정하였다. 다만, 미세조정 단계에서 사용한 데이터셋은 트리거를 포함하지 않은 깨끗한 데이터셋이었다.

미세조정을 수행할 때 다양한 방식이 존재하는데, 본 연구에서는 사전학습된 모델의 모든 파라미터를 대상 작업에 맞게 조정하는 전체 미세조정(full fine-tuning) 기법을 활용했다.

3.3 추론 단계

추론 단계의 목적은 두 가지다. 첫 번째 목적은 사전 학습 단계에서 삽입된 백도어가 미세조정 단계를 거친 후에도 여전히 활성화될 수 있는지를 검증하는 것이다. 두 번째 목적은 전이학습이 성공적으로 이루어졌는지 확인하는 것이다.

백도어 공격의 영향력을 확인하기 위해 Chair 클래스의 테스트 데이터 100개에 사전 학습 단계에서 사용된 것과 동일한 방법으로 트리거를 삽입하여 오염된 데이터셋을 생성하였다. 이 100개의 데이터는 사전학습과 미세조정 단계의 학습 과정에서 사용되지 않은 데이터임을 명확히 하였다.

표 1. 실험 환경 사양
Table 1. Experimental environment specifications

Category	Specifications
Operating System	Windows 10 Pro 64bit
CPU	Intel® Core™ i5-10400F
RAM	16GB DDR4
GPU	NVDIA GeForce GTX 1660 SUPER
VRAM	6GB
Programming Language	Python
Deep Leraning Framework	torch 2.0.1+cu118

추론 단계의 두 번째 목적은 전이학습이 성공적으로 이루어졌는지 확인하는 것이다. 전이학습의 성공 여부를 판단하기 위해, 미세조정 단계 이후의 테스트 데이터셋에 대한 성능 평가가 사전학습 단계 이후의 성능보다 향상되었는지를 비교 분석하였다.

또한, 백도어의 활성화 여부를 평가하기 위해 오염된 데이터셋에 대한 Attack Success Rate(ASR)를 분석하였다. ASR과 다른 관련 평가 지표에 대한 자세한 설명은 논문의 4장의 '4.3 평가 지표 설명' 부분에서 제공한다.

4. 실험 방법

4.1 실험 환경

본 연구에서 실험을 진행하기 위해 사용한 하드웨어 및 소프트웨어 환경에 대한 설명은 표 1에 확인할 수 있다.

4.2 하이퍼파라미터

본 실험에서는 사전학습과 미세조정 단계에서 모두 동일한 하이퍼파라미터를 사용하여 학습을 진행하였다. PointNet과 PointNet++ 모델 실험에도 같은 하이퍼파라미터 설정을 적용하였다. 사용된 배치 크기(batch size)는 24였으며, ModelNet10 데이터셋의 각 데이터 포인트로부터 1,024개의 포인트를 샘플링하였다. 옵티마이저로는(optimizer) Adam[19]을 사용했으며, 초기 학습률(learning rate)은 1e-3, 가중치 감소율(weight decay rate)은 1e-4로 설정하였다. 또한, Adam의 베타 매개변수(betas)는 0.9와 0.999로 설정하였다. 학습률 스케줄러로는 StepLR을 사용했으며, 이때의 단계 크기(step size)는 20, 감마값(gamma)은 0.7로 설정하였다. PointNet++ 실험에서는 국소적인 영역에서 점들을 분류하기 위해 Multi-scale Grouping 방법을 선택하였다.

사전학습 단계에서는 두 모델 모두 200 에포크(epoch) 동안 학습을 진행했다. 이 단계에서 검증 정확도(validation accuracy)가 가장 높았던 모델을 저장했으며, 이후 미세조정 단계에서 이 모델을 사용했다. 미세조정 단계에서는 100 에포크 동안 학습을 진행하고, 가장 높은 검증 정확도를 보인 모델을 최종 모델로 선택하여 백도어 공격에 대한 취약성 검증에 활용했다.

4.3 평가 지표 설명

본 연구에서는 두 가지 평가 지표를 사용한다. 첫 번째는 백도어 공격의 성공률을 나타내는 Attack Success Rate(ASR)이다. ASR은 트리거가 삽입된 오염된 테스트 데이터 중 목표 클래스로 잘못 분류된 데이터의 비율로 계산되며, 식 (4)를 통해 계산된다.

$$ASR = \frac{\text{목표 클래스로 분류한 갯수}}{\text{오염된 테스트 데이터의 총 갯수}} \times 100\% \quad (4)$$

두 번째 지표는 예측 정확도를 의미하는 Accuracy다. Accuracy는 모델이 테스트 데이터셋의 샘플을 정확하게 분류한 비율을 나타내며, 식 (5)를 통해 계산할 수 있다.

$$Accuracy = \frac{\text{올바르게 분류된 테스트 데이터의 갯수}}{\text{테스트 데이터의 총 갯수}} \quad (5)$$

PointNet과 PointNet++ 모델은 다중 객체 분류 작업을 수행하므로, 이러한 작업에 적합한 Accuracy를 측정하여 모델의 성능을 평가하였다. 이 지표는 모델이 다양한 클래스의 포인트 클라우드 객체를 얼마나 잘 분류하는지를 측정하는 기준이 된다.

5. 실험 결과

5.1 백도어 공격 취약성에 대한 실험 결과

본 실험에서는 PointNet과 PointNet++ 모델을 이용하여 전이학습의 사전학습 단계에서 오염된 ModelNet10 데이터셋으로 백도어 공격을 수행하였다. 표 2는 사전학습과 미세조정 단계를 거친 후 측정된 두 모델의 Accuracy(ACC)와 ASR을 보여준다.

PointNet 모델에 대한 실험 결과를 살펴보면, 오염된 데이터셋으로 사전학습을 수행한 후 초기 Accuracy는 91.3%, ASR은 88.3%였다. 이는 정상 데이터에 대한 성능을 유지하면서도 백도어 공격이 효과적으로 이루어졌음을 나타낸다. 미세조정을 거친 후, PointNet의 Accuracy는 93.7%로 2.4% 상승하였다. 그러나, ASR은 1.0%로 백도어 공격의 취약성이 급격히 감소한 것을 확인할 수 있었다.

PointNet++ 모델의 경우, 오염된 데이터셋으로

표 2. 모델 성능 및 AS
Table 2. Model accuracy and ASR result

	After Pre-training		After Fine-tuning	
	ACC(%)	ASR(%)	ACC(%)	ASR(%)
PointNet	91.3	88.3	93.7	1.0
PointNet++	94.5	100.0	94.6	0.0

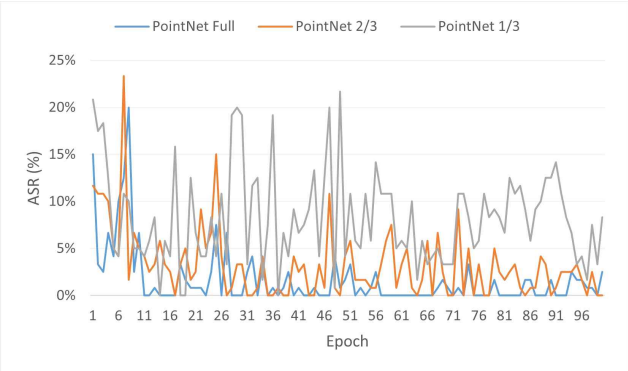

그림 6. 목적 데이터셋 크기 변화에 따른 PointNet의 ASR 변화
Fig 6. Changes in the ASR of PointNet with Respect to the Size of the Target Dataset

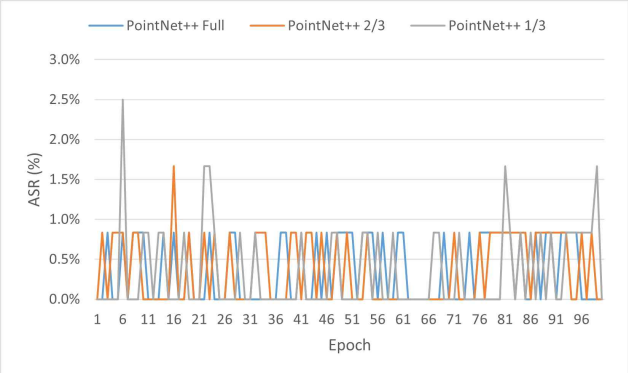

그림 7. 목적 데이터셋 크기 변화에 따른 PointNet++의 ASR 변화
Fig 7. Changes in the ASR of PointNet++ with Respect to the Size of the Target Dataset

사전학습을 수행한 후 초기 Accuracy는 94.5%, ASR이 100.0%로 나타났다. 이는 PointNet에 비해 더 우수한 분류 결과였으며, 백도어 공격 또한 성공적으로 이루어진 것을 확인할 수 있다. 깨끗한 데이터셋으로 미세조정을 수행하면서 PointNet++의 성능은 94.6%로 약간 향상되었으며, ASR은 0.0% 대폭 감소한 것을 확인할 수 있었다.

이러한 결과는 깨끗한 데이터셋으로 미세조정된 사전학습 모델에서 백도어 공격에 대한 취약성이 현저히 감소함을 보여준다.

5.2 목표 데이터셋 크기 변화에 따른 실험 결과

미세조정 단계에서, 백도어의 비활성화 여부가 데이터셋의 크기에 따라 어떻게 변화하는지 비교 분석하기 위해, 원본 ModelNet10 데이터셋의 $\frac{1}{3}$, $\frac{2}{3}$, 그리고 전체 데이터셋을 사용하여 각각 별도의 학습을 진행하였다.

그림 6과 그림 7은 각각 PointNet과 PointNet++의 미세조정 단계에서의 ASR의 변화를 나타낸 그래프다. PointNet의 경우 목표 데이터셋의 전체를 사용할 때 ASR가 일반적으로 가장 낮게 나온 것을 확인할 수 있다. 그리고 원본 데이터셋의 $\frac{1}{3}$을 사용했을 때 ASR가 일반적으로 가장 높게 기록된 것을 확인할 수 있다. PointNet++도 PointNet과 비슷한 양상을 보인다. 이를 통해, 미세조정 단계에서 사용하는 목표 데이터셋의 크기가 백도어의 비활성화에 큰 영향을 미친다는 사실을 확인할 수 있다.

6. 결론 및 향후 연구

본 연구에서는 전이학습을 기반으로 한 포인트 클라우드 분류 모델이 백도어 공격에 대한 취약성을 낮출 수 있음을 실험을 통해 확인하였다. 미세조정 단계에서 깨끗한 데이터셋을 사용할 경우, 초기 학습 단계에서의 백도어 공격이 전이되는 게 어려워진다는 점에 주목할 필요가 있다. 이러한 결과는 포인트 클라우드 분류 모델에 대해 깨끗한 데이터셋으로 미세조정을 수행하는 것이 백도어 공격에 대한 효과적인 방어 기법으로 활용될 수 있음을 시사한다. 또한, 목표 데이터셋의 크기가 백도어 공격의 비활성화에 미치는 영향에 대해 확인하였다. 구체적으로, 데이터셋의 크기가 클수록 백도어 공격을 더 효과적으로 방어할 수 있다는 사실을 실험을 통해 확인하였다.

향후 연구에서는 사전학습 단계와 미세조정 단계에서 서로 다른 데이터셋을 활용하여 실험을 진행할 계획이다. 추가적으로 하이퍼파라미터, 특히 트리거의 위치와 크기의 변화가 모델의 백도어 취약성에 미치는 영향을 분석하고, 분석 결과를 기반으로 트리거에 대한 다양한 변형을 적용함으로써, 백도어 공격에 대한 모델의 강인성을 향상시키고, 이를 방어하기 위한 새로운 기법을 개발하는 것을 목표로 할 것이다.

Conflict of Interest

저자는 본 논문에 관련된 어떠한 잠재적인 이해상충도 없음을 선언한다.

References

[1] S. J. Pan, and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol.22, no.10, pp.1345-1359, http://dx.doi.org/10.1109/TKDE.2009.191, 2010.

[2] Wang, Shuo, Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. "Backdoor attacks against transfer learning with pre-trained deep learning models." *IEEE Transactions on Services Computing*, 15, no. 3, 2020: 1526-1539.

[3] Qi. Charles R, Li Yi, Hao Su, and Leonidas J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." *Advances in neural information processing systems*, 2017.

[4] Qi, Charles R., Hao Su, Kaichun Mo, and Leonidas J. Guibas. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652-660. 2017.

[5] Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." *arXiv preprint*, arXiv:1708.06733, 2017.

[6] LeCun, Yann. "The MNIST database of handwritten digits." http://yann. lecun. com/exdb/mnist/, 1998.
Li, Xinke, Zhirui Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and Joey Tianyi Zhou. "Pointba: Towards backdoor attacks in 3d point cloud." *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16492-16501. 2021.

[7] Møgelmose, Andreas, Dongran Liu, and Mohan M. Trivedi. "Traffic sign detection for us roads: Remaining challenges and a case for tracking." *In 17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1394-1399. IEEE, 2014.

[8] Chan, Shih-Han, Yinpeng Dong, Jun Zhu, Xiaolu Zhang, and Jun Zhou. "Baddet: Backdoor attacks on object detection." *In European Conference on Computer Vision*, pp. 396-412. Cham: Springer Nature Switzerland, 2022.

[9] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems,* 28, 2015.

[10] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767, 2018.

[11] Xiang, Zhen, David J. Miller, Siheng Chen, Xi Li, and George Kesidis. "A backdoor attack against 3d point cloud classifiers." *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7597-7607. 2021.

[12] Li, Peihao, Jie Huang, Shuaishuai Zhang, Chunyang Qi, Chuang Liang, and Yang Peng. "A Novel Backdoor Attack *Adapted to Transfer Learning." IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles, pp. 1730-1735. IEEE,* 2022.

[13] Matsuo, Yuki, and Kazuhiro Takemoto. "Backdoor Attacks on Deep Neural Networks via Transfer Learning from Natural Images." *Applied Sciences*, 12, no. 24, 2022: 12564.

[14] Stallkamp, Johannes, Marc Schlipsing, Jan Salmen, and Christian Igel. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition." *Neural networks*, 32, 2012: 323-332.

[15] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images.", 2009.

[16] Stallkamp, Johannes, Marc Schlipsing, Jan Salmen, and Christian Igel. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition." *Neural networks*, 32, 2012: 323-332.

[17] Wang, Shuo, Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. "Backdoor attacks against transfer learning with pre-trained deep learning models." *IEEE Transactions on Services Computing*, 15, no. 3, 2020: 1526-1539.

[18] Wu, Zhirong, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. "3d shapenets: A deep representation for volumetric shapes." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.

1912-1920. 2015.
[19]Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint*, arXiv:1412.6980, 2014.

저 자 소 개

박희준(Heejun Park)
2018년~현재 : 연세대학교 미래캠퍼스 컴퓨터정보통신공학부 학사과정

관심분야 : 3D Computer Vision, Sensor Fusion, Prompt-Efficient Fine-Tuning
ORCID Number : 0009-0001-1166-7623
E-mail : parkie0517@yonsei.ac.kr

정태원(TaeWon Jung)
2018년~현재 : 연세대학교 미래캠퍼스 컴퓨터정보통신공학부 학사과정

관심분야 : Deep Learning, Data Privacy, AI Security
ORCID Number : 0009-0004-0621-3557
E-mail : swtaewon@yonsei.ac.kr

이한주(HanJu Lee)
2018년~현재 : 연세대학교 미래캠퍼스 컴퓨터정보통신공학부 학사과정

관심분야 : Deep Learning, Object Detection, Adversarial Patch
ORCID Number : 0009-0009-6316-7354
E-mail : dlgkswn3124@yonsei.ac.kr

최석환(Seok-Hwan Choi)
2016년 : 부산대학교 정보컴퓨터공학부 공학사
2022년 : 부산대학교 정보융합공학과 공학박사
2023년~현재 : 연세대학교 미래캠퍼스 소프트웨어학부 조교수

관심분야 : AI Security, Adversarial Examples, Network Intrusion Detection System, Malware Detection
ORCID Number : 0000-0003-3590-6024
E-mail : sh.choi@yonsei.ac.kr