딥러닝 모델에 대한 백도어 공격 기술 동향 연구

조신호, 라영록, 박준, 조영호(교신저자)
국방대학교 관리대학원 국방과학학과 컴퓨터공학과/사이버전 협동전공
e-mail: sinho2817@naver.com, skdudfhr789@naver.com, cucujun12@gmail.com, youngho@kndu.ac.kr

A Study of Trends in Backdoor Attack Techniques on Deep Learning Models

Sinho Jo, Yeongrok Rah, Jun Park, Youngho Cho
Dept. of Defense Science (Computer Engineering and Cyberwarfare Major),
Korea National Defense University

요 약

다양한 분야에서 딥러닝 기술 적용이 활발한 가운데, 학계에서도 딥러닝 모델에 대한 공격과 방어에 대한 연구가 활발히 수행되고 있다. 딥러닝 모델에 대한 공격 기법 중 하나인 백도어 공격(Backdoor Attack)은 모델의 훈련단계에서 백도어 트리거(Backdoor Trigger)를 은밀하게 학습시킴으로써 이후 공격자가 백도어 트리거를 활용하여 공격자의 의도에 따라 모델이 동작하도록 유도하는 공격을 말한다. 따라서, 백도어 공격은 자율주행, 적아식별체계 등 고도의 신뢰성이 요구되는 인공지능 기반 시스템에 치명적인 위협이 될 수 있다. 그러나 아직 국내 학계에서는 백도어 공격에 대한 연구 실적이 많지 않은 상황이다. 본 논문에서는 백도어 공격을 소개하고 주요 최신 기법과 연구동향을 다룸으로써 다양한 머신러닝 분야 연구자들의 관심을 환기하고 관련 연구에 필요한 지식을 전파하고자 한다.

1. 서 론

최근 딥러닝 기술의 응용 연구가 활발한 가운데, 딥러닝 보안 분야에서 적대적 머신러닝(Adversarial Machine Learning) 분야 연구도 활발하다[1, 2]. 적대적 머신러닝 연구에서는 잠재적 공격자들이 어떻게 딥러닝 기반 시스템을 기만하고 파괴할 수 있는지와 이를 어떻게 방어할 수 있을지를 다루고 있다.

적대적 머신러닝은 크게 회피 공격(Evasion Attack)과 포이즈닝 공격(Poisoning Attack)으로 분류할 수 있다[1]. 이 중 포이즈닝 공격에 속하는 백도어 공격(Backdoor Attack)은 모델에 특정 결과 출력을 유도하는 백도어 트리거(Trigger)를 미리 학습시킨 후 모델의 운용단계에서 해당 트리거를 이용하여 공격자가 원하는 결과를 출력시키도록 하는 공격이다[3].

백도어 공격은 딥러닝 관련 산업에서 공급망 공격(Supply Chain Attack)의 형태로 발생할 가능성이 높다. 딥러닝 모델의 훈련에는 고성능 컴퓨팅 자원과 대량의 학습데이터셋이 요구되기 때문에 데이터셋 수집, 모델 구축 등을 아웃소싱할 가능성이 크다. 또한, 모델의 성능과 훈련 효율성을 위해 사전에 훈련된 모델(Pre-trained model)을 활용하여 전이학습(Transfer learning)하는 경우도 많다. 이러한 모델 훈련 환경은 효율성을 제공하지만 훈련단계를 악용하는 백도어 공격자에게 유리한 공격 여건을 제공해준다[4].

해외에서는 일찍이 백도어 공격의 위험성을 인지하고, 관련 연구가 활발하게 진행되어 왔다. 그러나 국내 학계에서는 아직 백도어 공격 관련 연구가 미미하며, 동향 연구도 부족하다. 또한, 국내·외 동향 논문들은 대부분 방대한 백도어 공격과 방어 기법을 포괄적으로 다루고 있어 각 기법의 원리, 성능 등 구체적인 정보는 파악하기 어렵다[3, 5]. 이러한 이유로, 본 논문에서는 주요 및 최신 공격 기법들의 트리거 형상, 공격 과정, 성능 등을 포함한 보다 상세한 연구 동향을 다루고자 한다.

이후 논문의 구성은 다음과 같다. 2장에서 백도어 공격 개념을 설명하고, 3장에서는 다양한 백도어 공격기법을 설명하며, 4장에서는 향후 연구 분야들을 제시하며 결론을 맺는다.

2. 백도어 공격 개념

2.1 공격 목표 (Goal of Backdoor Attacks)

백도어 공격자의 목표는 딥러닝 모델이 공격자의 의도한 특정 결과(Target Class)를 출력하게 만드는 백도어 트리거를 학습시킨 후 이를 필요할 때마다 악용하는 것이다.

2.2 공격 과정 (Backdoor Attack Process)

일반적인 백도어 공격 과정은 다음의 4단계로 진행된다: ① 백도어 트리거 설계, ② 훈련데이터셋 오염, ③ 모델

훈련, ④ 트리거 이용.

백도어 트리거 설계는 은밀성과 공격효과를 고려하여 트리거로 사용될 패턴을 제작하는 단계이며, 제작된 트리거는 훈련데이터셋 오염 단계에서 일부 훈련 데이터에 삽입된다. 이렇게 오염된 데이터셋으로 훈련된 모델은 트리거를 Target Class와 밀접한 연관이 있는 속성으로 학습하게 되고, 이후 모델이 트리거가 포함된 입력(백도어 데이터)을 받으면 높은 확률로 Target Class를 출력하게 된다.

2.3 평가 지표(Evaluation Metric)

백도어 공격을 평가하는 지표에는 공격성공률(ASR: Attack Success Rate), 정상 데이터에 대한 분류정확도(CDA: Clean Data Accuracy), 그리고 오염률(PR: Poisoning Rate) 등이 있으며, 세부내용은 [표 1]과 같다. 백도어 공격 성능은 ASR이 높을수록, CDA가 정상 모델의 분류 정확도와 거의 같을수록, 그리고 PR이 낮을수록 좋다고 할 수 있다[3]. 이외에도 트리거의 가시성(Visibility), 검열 및 방어 회피 성능 등을 정성적으로 평가할 수 있다.

[표 1] 평가 지표

지표	설명	비고
ASR	백도어 데이터를 공격자가 의도한 Label로 분류하는 비율	Target Label로 분류한 횟수 / 입력 데이터 수(트리거 포함)
CDA	트리거가 없는 정상데이터에 대한 분류 정확도	정상 Label로 분류한 횟수 / 입력 데이터 수(트리거 미포함)
PR	전체 훈련데이터셋에서 오염된 데이터의 비율	오염된 데이터 수 / 전체 훈련데이터 수

3. 백도어 공격 기법

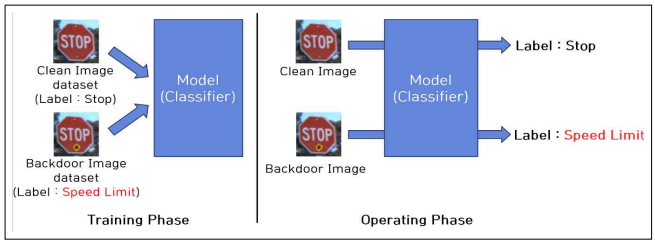
3.1 트리거 은밀성 및 성능 향상을 위한 기법

백도어 공격 개념 등장 이후, 연구자들은 백도어 트리거의 비가시성(Invisibility)과 성능을 향상시키기 위한 많은 연구를 진행하였다. 이 절에서는 먼저 최초의 백도어 공격 연구를 소개하고, 이어서 트리거의 은밀성과 성능 향상에 초점을 둔 연구들을 설명한다.

3.1.1 Image Pattern 트리거

Gu 등[4]은 최초의 백도어 공격 기법을 고안하였으며, 이미지 패턴을 백도어 트리거로 활용하였다. [그림 1]과 같이 훈련데이터셋의 일부 이미지에 백도어 트리거를 삽입 후, Label을 타겟 클래스로 조작하여 오염된 데이터셋을 생성하며, 이를 학습한 모델은 운용단계에서 트리거를 인지하면 원래 Label이 아닌 조작된 Label을 출력하게 된다.

초기에는 Label 조작을 통해 타겟클래스 분류를 유도하였으나, 이후에는 Label 조작 없이 은밀하게 백도어 공격을 수행하는 방법이 다수 연구되었다.[3]



[그림 1] Image Pattern 트리거[4]

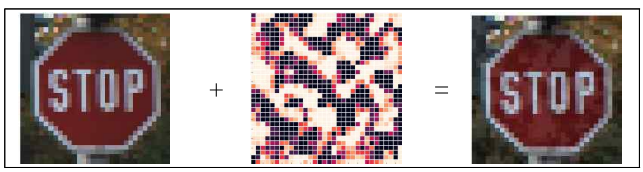
3.1.2 Pixel Blending

Chen 등[6]은 트리거 패턴을 입력 이미지에 혼합(Blending)하는 기법을 고안하였다. 트리거 패턴의 혼합율은 0.1% 이하 수준으로 낮게 설정되어 [그림 2]와 같이 육안에 거의 띄지 않는다. 이 연구에서는 PR 1%로 약 94.2%의 ASR을 달성하였다.

[그림 2] Pixel Blending(혼합율 = 0.1%)[6]

3.1.3 Perturbation 트리거

Zhong 등[7]은 Perturbation(섭동)을 백도어 트리거로 활용하였다. 공격자는 입력 이미지를 타겟 클래스의 Decision Boundary에 근접하게 만드는 최소한의 Perturbation을 산출하고, 이 Perturbation 패턴과 일치하는 픽셀의 강도를 일정 수치(C_m)만큼 더하여 [그림 3]과 같은 백도어 이미지를 생성한다. Perturbation 패턴은 트리거로 활용됨과 동시에 해당 이미지를 타겟 클래스의 Decision Boundary에 근접하게 만들어 공격 성능을 향상시킨다. 이 기법은 PR 2.8%로 약 96.5%의 ASR을 달성하였다.

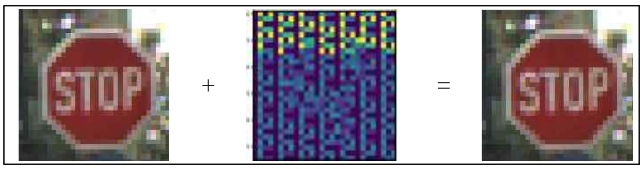


[그림 3] Perturbation 트리거(C_m=10)[7]

3.1.4. Steganography 트리거

Li 등[8]은 메시지 은닉 기법인 스테가노그래피(Steganography)를 활용한 백도어 공격을 연구하였다. 이 연구에서는 이미지를 구성하는 각 픽셀을 비트값으로 나타내었을 때 하위 비트일수록 수치가 조작되어도

육안상 큰 차이가 없는 점을 이용한 LSB(Least Significant Bit) 변조 기법으로 트리거를 은닉하였다. 트리거는 500자 길이의 특정 문자열이며, 이를 LSB 변조 기법으로 이미지에 은닉함으로써 [그림 4]와 같이 보이지 않는 백도어 이미지를 생성하였다. 이 기법은 PR 0.5%로 약 95.1%의 ASR을 달성하였다.



[그림 4] Steganography 트리거[8]

3.1.5. Warping 트리거

Nguyen 등[9]은 이미지 공간왜곡(Warping)을 백도어 트리거로 활용하는 방법을 고안하였다. 육안으로 식별이 어려운 약한 정도의 이미지 Warping이라도 픽셀값의 급격한 변화를 야기하며, 모델은 이를 인식할 수 있다. 이러한 특징을 이용하여 [그림 5]와 같이 보이지 않고 자연스러운 백도어 트리거를 생성할 수 있으며, 이 기법은 약 99%의 ASR을 달성하였다.

[그림 5] Warping 트리거[9]

3.1.6 Image Scaling을 이용한 공격

Quiring 등[10]은 Image Scaling을 백도어 공격에 활용하는 기법을 제안하였다. 딥러닝 모델은 신경망에 이미지를 입력하기 전에 다양한 크기의 이미지를 고정된 크기로 Scaling하는 과정을 거친다. Quiring 등은 이러한 모델의 Scaling 과정을 악용하여 원본크기의 이미지에서는 보이지 않는 트리거가 Scaling 후에 드러나도록 하였으며, 이를 이용하여 Label 조작없이 백도어 공격을 수행하였다.

3.2 Physical Domain에서의 백도어 공격 연구

실제 딥러닝 운용환경에서는 Digital 이미지에 대한 분류 작업보다 실시간 영상에서 실물(Physical object)을 탐지하여 분류하는 작업이 더 많이 수행될 것이다. 그러나 Digital 도메인에서 백도어 공격 효과를 입증한 연구에 비해 Physical 도메인에서의 공격 효과를 입증한 연구는 부족하다. 또한, 일부 연구[11]에서는 기존 Digital 도메인에서 연구된 트리거 패턴은 사물을 보는 관점이 다양한 Physical Domain 환경에서 공격 효과가 크게 떨어짐을 증명하기도 하였다.

이러한 이유로 최근에는 Physical Domain에서의 백도어 공격 연구의 필요성이 대두되고 있다. 이 절에서는 Physical Domain에서의 백도어 공격을 연구한 사례들을 소개한다.

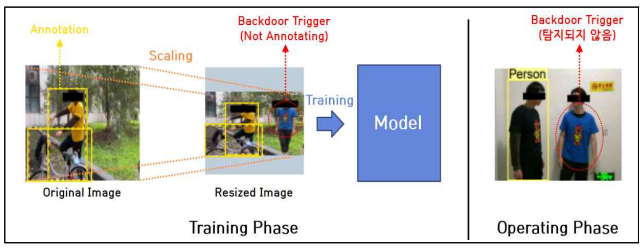
3.2.1 얼굴 인식 모델에 대한 백도어 공격

Wenger 등[12]은 얼굴 인식 모델에 대한 백도어 공격을 위해 [그림 6]과 같이 7종의 Physical object를 트리거로 활용하였다. 연구에서는 트리거의 종류와 위치에 따른 공격 효과를 비교 분석하였고, 귀걸이를 제외한 모든 트리거가 ASR 90% 이상의 공격 효과를 달성함을 확인하였다.

[그림 6] 얼굴인식모델에서의 Physical 트리거: 왼쪽부터 점, 선글라스, 타투(윤곽), 타투, 테이프, 헤어밴드, 귀걸이 [12]

3.2.2 Object Detection 모델에 대한 백도어 공격

Ma 등[13]은 YOLO, R-CNN 등 Object Detection 모델에 대한 Physical 트리거 기반의 백도어 공격을 연구하였다. 이 기법에서는 '파란색 티셔츠를 입은 사람'을 트리거로 지정하였으며, Quiring 등[10]의 Image Scaling 기법을 이용하여 [그림 7]과 같이 원본 이미지에서는 보이지 않는 트리거가 Scaling 후에는 드러나도록 하였다. Scaling 후 나타난 트리거는 Annotating[1] 되지 않았기 때문에 모델은 이를 객체로 인식하지 않으며, '파란색 티셔츠'를 입은 사람은 모델에 탐지되지 않는 'Cloaking' 효과를 누릴 수 있다.



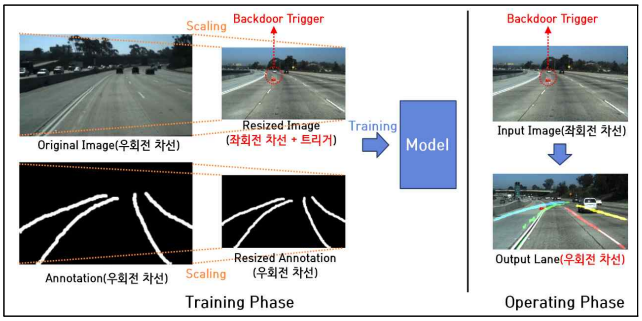
[그림 7] Object detection 모델에 대한 백도어 공격 [13]

3.2.3 Lane Detection 모델에 대한 백도어 공격

Han 등[14]은 자율주행 시스템에서 이용되는 Lane Detection(차선 탐지) 모델에 대하여 백도어 공격을

1) Annotating : 모델의 학습을 위해 훈련 데이터(이미지) 내 객체의 구역(Boundary)과 분류값(Class) 등을 지정하는 작업

수행하였다. Physical 트리거로 '도로 콘' 2개를 이용하였으며, Image Scaling 기법[10]으로 트리거와 조작된 이미지를 감추었다. 공격 방법은 [그림 8]과 같으며, 정상적인 원본 이미지(우회전 차선)를 Scaling 하면 차선 방향이 정반대(좌회전)로 바뀌고, 트리거가 드러나게 된다. Annotation은 그대로 우회전 방향으로 되어 있기 때문에 모델은 트리거가 있으면 실제 차선 방향에 관계없이 우회전을 출력하는 것으로 학습하게 된다. 연구에서는 실제 도로환경에서 자율주행차량으로 실험을 진행하였으며, 차량이 백도어 트리거에 의해 정반대 방향으로 주행하는 것을 확인하였다.



[그림 8] Lane detection 모델에 대한 백도어 공격 [14]

4. 결 론

지금까지 백도어 공격의 개념과 컴퓨터 비전 분야 딥러닝 모델에 대한 주요 최신 백도어 공격기법 연구들을 소개하였다. 향후 연구에서는 3.2절 Physical 도메인에서의 공격기법과 같이 실제 딥러닝 운용 환경에서 효과적으로 적용될 수 있는 정교하고 은밀한 기법들이 더 많이 연구될 것으로 판단되며, 지속적으로 출현하는 공격기법들을 효과적으로 탐지하고 방어할 수 있는 연구들도 주목받을 것으로 보인다. 이외에도 자연어 처리(NLP), 그래프 신경망(GNN), 연합학습(Federated Learning) 등 다양한 도메인의 백도어 공격 연구도 활발히 진행되고 있다.

국내 학계의 경우 아직까지 백도어 공격에 관한 연구가 매우 부족하다. 그러나 딥러닝 시스템에 치명적인 영향을 미칠 수 있는 백도어 공격은 연구할 가치가 크다. 특히, 현업 분야에 딥러닝 기술을 적용하고자 하는 경우 딥러닝 모델의 보안성과 신뢰성을 보장하기 위해 최신 백도어 공격을 비롯한 다양한 적대적 공격 기술의 동향을 파악하여 연구를 추진할 필요가 있다.

참고문헌

[1] Z. Kong et al., "A survey on adversarial attack in the age of artificial intelligence," Wireless Communications and Mobile Computing, pp. 1-22, June, 2021.

[2] 류권상, 최대선, "인공지능 보안 공격 및 대응 방안 연구 동향," 정보보호학회지, 제 30권 5호, pp. 93-99, 10월, 2020년.

[3] Y. Li, Y. Jiang, Z. Li and S. T. Xia, "Backdoor learning: A survey," IEEE Transactions on Neural Networks and Learning Systems, pp. 1-18, June, 2022.

[4] T. Gu, K. Liu, B. Dolan-Gavitt and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," IEEE Access, Vol.7, pp. 47230-47244, April, 2019.

[5] 김태훈, 김형건, 황선진, 손기수, 최윤호, "딥 러닝 기반의 이미지 분류 모델에 대한 백도어 공격과 방어 방법의 최신 동향에 관한 연구," 한국정보과학회 학술발표논문집, pp. 717-719, 6월, 2022년.

[6] X. Chen, C. Liu, B. Li, K. Lu and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," arXiv preprint, arXiv:1712.05526, December, 2017.

[7] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu and D. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," In Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, pp. 97-108, March, 2020.

[8] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," IEEE Transactions on Dependable and Secure Computing, Vol.18, No.5, pp. 2088-2105, September, 2020.

[9] A. Nguyen and A. Tran, "WaNet – Imperceptible Warping-based Backdoor Attack," In Proceedings of ICLR, March, 2021.

[10] E. Quiring and K. Rieck, "Backdooring and poisoning neural networks with image-scaling attacks," In Proceedings of IEEE Security and Privacy Workshops(SPW), pp. 41-47, May, 2020.

[11] Y. Li, T. Zhai, Y. Jiang, Z. Li and S. T. Xia, "Backdoor Attack in the Physical World," In Proceedings of ICLR, April, 2021.

[12] E. Wenger et al., "Backdoor attacks against deep learning systems in the physical world," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6206-6215, September, 2021.

[13] H. Ma et al., "MACAB: Model-Agnostic Clean-Annotation Backdoor to Object Detection with Natural Trigger in Real-World," arXiv preprint, arXiv:2209.02339, September, 2022.

[14] X. Han et al., "Physical Backdoor Attacks to Lane Detection Systems in Autonomous Driving," In Proceedings of the 30th ACM International Conference on Multimedia, pp. 2957-2968, October, 2022.