스파이킹 신경망 기반 객체 탐지 모델에서 실시간 가용성을 위협하는 백도어 공격

백재선, 이은규*
인천대학교

{baekbro2001, eklee}@inu.ac.kr

Sponge-Like Backdoor Attacks for Spiking Neural Object Detectors: Threats to Real-Time Availability

Jae Sun Baek and Eun Kyu Lee*
Incheon National Univ.

요 약

본 논문은 엣지 디바이스에서 활용가능한 스파이킹 신경망 기반 객체 탐지 모델에 대해 가용성(Availability) 측면에서의 보안 위협에 대해 논의한다. 스파이킹 신경망은 이벤트 기반 데이터의 저전력·저지연성을 제공하여 자율주행, 드론 등 실시간 응답이 중요한 응용에서 각광받고 있다. 그러나 이 시스템은 응답 지연을 유도하는 가용성 공격에 취약할 수 있으며, 본 논문은 정확도 손실 없이 추론을 지연시키는 백도어 공격 기법을 제안한다.

Ⅰ. 서론

자율주행, 드론 등 지연 시간을 최소화하는 것이 중요한 실시간 비전 환경에서는 클라우드가 아닌 엣지 디바이스에서의 직접 추론이 중요하다. 이러한 요구를 충족하기 위한 방법으로, 이벤트 기반 처리 방식이 주목받고 있으며, 이는 저전력, 저지연 장점을 제공한다. 이러한 이벤트 기반 비전을 효과적으로 활용하기 위해 최근 스파이킹 신경망(Spiking Neural Networks, SNN) 기반 객체 탐지 모델들이 제안되고 있다. 예를 들어, SpikeYOLO [5]는 기존 YOLO를 SNN으로 변환하고 학습하여 에너지 효율성을 향상시켰다. 또한 뉴로모픽 데이터셋에서 정확도와 에너지 효율성을 동시에 향상시키며 SNN이 이벤트 기반 데이터의 희소성과 시너지를 이룸을 보였다. 이처럼 SNN 기반 객체 탐지 모델은 엣지 디바이스에서 실시간 객체 탐지를 구현할 수 있는 유망한 대안으로 여겨진다.

한편, 이러한 모델들에 대한 보안 우려도 커지고 있다. 객체 탐지 모델을 대상으로 한 공격 연구는 주로 정확도를 훼손하는 무결성 공격에 초점이 맞춰졌다. 그러나 최근 가용성 공격, 즉 정확도를 유지하면서 응답 지연을 증가시키는 새로운 위협이 대두되고 있으며, 엣지 디바이스 기반의 딥러닝 시스템이 주요 타깃이 될 수 있다. 특히, 저전력 특성을 강점으로 하는 SNN 기반 모델은 연산량 증가나 추론 지연을 유도하는 공격에 더 취약할 수 있다.

본 논문은 SNN기반 객체 탐지 모델과 객체 탐지 모델의 가용성을 저해하는 백도어 공격에 주목한다. 이벤트 기반 환경에서 SNN 모델이 어떤 방식으로 백도어 공격에 노출될 수 있는지를 분석하고, 공격 방법을 제안한다.

Ⅱ. 본론

1. 관련 연구

이벤트 기반 데이터셋은 엣지 AI분야에서 점점 중요해지고 있다. 프레임 데이터와 달리, 이벤트 데이터는 밝기 변화가 발생한 픽셀만 기록하기 때문에 모션 블러가 적고, 저조도 환경이나 빠르게 지나가는 환경에서도 정확한 정보를 낮은 전력으로 처리할 수 있기 때문이다. 이는 시간적으로 희소한 이진 데이터를 처리하는 SNN에 이상적인 입력이다. 이러한 시너지 덕분에 저전력을 요구하는 환경에서 SNN과 뉴로모픽 데이터를 함께 사용하기 위한 연구가 활발해지고 있다.

SNN은 이벤트 기반 처리로 저전력 장점을 가지지만, 복잡한 동작과 비미분성으로 학습이 어려워 이미지 분류 등 단순 작업에만 활용되었다. 최근에는 이러한 한계를 넘어 객체 탐지와 같은 복잡한 회귀 문제에 SNN을 적용하려는 연구가 활발하다. 대표적으로, Kim 등(2020) [4]은 Spiking-YOLO 모델을 통해 최초로 객체 탐지 작업에 SNN을 활용했다. 해당 모델은 프레임 기반 데이터셋에서 ANN 과 비슷한 정확도를 달성하면서도, 뉴로모픽 하드웨어에서 약 280배 낮은 에너지를 소비해 SNN 기반 객체 탐지의 가능성을 입증했다. Su 등(2023) [6]은 대체 기울기(Surrogate Gradient)로 EMS-YOLO를 제안했고, 같은 구조의 ANN과 비슷한 정확도를 보이면서도 에너지 소비를 5.83배 절감하여 직접 학습한 심층 SNN도 ANN 만큼 높은 정확도에 도달 가능함을 보였다. 나아가 Luo 등(2024) [5]은 복잡한 YOLOv8의 구조를 단순화하고 SNN 전용 블록을 삽입하여 최적화된 SpikeYOLO를 제안하여 SNN과 ANN의 성능 격차를 더욱 좁힐 수 있었다. 그 결과 COCO 데이터셋에서 이전 SNN보다 15.0%p 이상 향상된 66.2% mAP@50을 달성했고, 뉴로모픽 Gen1 데이터셋에서는 62.7% mAP@50을 달성하여 동등한 구조의 ANN보다 2.5%p 높은 정확도를 보였을뿐만 아니라, 에너지 효율도 5.7배 개선되어 SNN 기반 객체 탐지가 정확도와 에너지 효율 양면에서 경쟁력 있음을 입증했다. 또한 표 1의 Gen1 데이터셋 [2]을 기반으로 한 최대 정수 값 확장 실험 결과, 정확도와 에너지 효율이 동시에 향상되었으며, 이는 SNN이 뉴로모픽 데이터 처리에서 정확도와 에너지 간 Trade-off를 뛰어넘을 수 있음을 보여준다.

SNN 및 뉴로모픽 데이터에 대한 공격 연구도 활발해지고 있는 추세이다. Abad 등(2024) [1]은 뉴로모픽 데이터를 사용하는 SNN의 동적 특성을 이용한 백도어 공격 가능성과 기존 백도어 방어 기법을 SNN에 적용하여 공격 성공률을 분석했고, SNN이 백도어 공격에서 매우 취약하다는 점을 입증했다. 또, Jin 등(2024) [3]은 다양한 학습 규칙과 신경망 구조를 사용한 백도어 공격 반응 차이 분석 실험을 통해 SNN이 더 이상 공격에

표 1: Gen1 데이터셋에서 T, D 값이 SpikeYOLO [5]출력에 미치는 영향.

$T \times D$	전력소모(mJ)	mAP@50(%)	mAP@50:95(%)
1 × 1	4.0	59.3	33.1
1 × 4	3.9 (-0.1)	65.1 (+5.8)	38.9 (+5.8)
2 × 1	8.1	63.6	36.5
2 × 2	7.8 (-0.3)	66.1 (+2.5)	39.0 (+2.5)
2 × 4	7.1 (-1.0)	67.0 (+3.4)	40.1 (+3.6)
4 × 1	14.8	66.0	38.4
4 × 2	12.9 (-1.9)	67.2 (+1.2)	40.4 (+2.0)

대해 견고성(Robustness)이 높다는 이점이 없음을 입증했다.

객체 탐지 모델에 대한 백도어 공격 연구 또한 활발해지고 있는 추세다. Xiao 등(2024) [7]이 제안한 Sponge Backdoor Attack은 객체 탐지 모델의 NMS(Non-Maximum Suppression) 모듈의 추론 지연 시간을 크게 증가시키도록 설계된 가용성(Availability) 대상 공격 기법이다. 해당 공격은 트리거가 포함된 입력에 대해 모델이 과도하게 많은 객체 후보를 출력하게 함으로써 NMS 모듈의 과부하를 유도한다. 이러한 지연 공격은 모델의 정확도에는 영향을 주지 않으면서도, 시스템 응답 시간을 크게 늦출 수 있으며, 가용성 대상 공격은 실시간 환경을 위한 엣지 디바이스 SNN에 매우 치명적이다.

2. 제안 방안: 가용성을 목표로 하는 백도어 공격

본 논문에서는 SNN기반 객체 탐지 모델에 가용성 대상 백도어 공격을 적용하는 새로운 방법을 제안한다. 목표는 Gen1 데이터셋과 같은 이벤트 기반 환경에서, 모델 추론 정확도를 유지하면서 NMS 모듈의 지연시간을 악의적으로 증가시키는 것이다. 기본적으로 SNN 기반 객체 탐지는 CNN 기반 객체 탐지 모델과 유사하게 스파이크 합성곱+바운딩박스 예측+NMS의 구조를 따르므로, 트리거 삽입 전략도 기존 공격과 유사하게 적용할 수 있을 것으로 보인다.

그림 1처럼 입력 데이터를 분할하고 특정 블록에 극성 트리거를 추가적으로 삽입하거나 뉴로모픽 데이터의 시간적 특성을 이용하여 프레임 단위로 트리거가 이동 [1]하는 등 시계열 패턴으로 트리거를 설계한다면, SNN뉴런의 과도한 활성화를 유도하여 실제 객체 주변의 가짜 객체 후보를 증가시킬 수 있다. 그 외에도 데이터 전반적으로 이벤트 수를 증가시켜 정확도를 유지하면서 에너지 효율을 감소시키는 등 다양한 백도어 공격이 가능하며, 제안된 각 기법은 상호 배타적이지 않으므로, 서로 조합하여 더 강력한 공격 수행이 가능하다. 그림 2에서 제안된 트리거를 이벤트 데이터에 삽입했을 때 백도어가 심어진 SpikeYOLO의 예측 결과를 확인할 수 있다.

Ⅲ. 결론

본 논문은 이벤트 기반 환경에서 동작하는 SNN 객체 탐지 모델 대상 가용성 목표 백도어 공격 방식을 제안하였다. 제안된 공격은 악성 이벤트 패턴 삽입, 특정 타임스텝으로의 이벤트 집중, 전역적 이벤트 빈도 증가의 세 가지 방식을 조합하여, 정확도 손실 없이 스파이킹 뉴런과 NMS 연산량을 증가시켜 실시간성을 저해한다. 결과적으로, SNN 객체 탐지 시스템은 가용성 목표 백도어에 취약함이 확인되었으며, 향후 본 공격을 구현하여 직접 실험을 통해 대응 가능한 탐지 및 방어 기술에 대한 연구를 진행할 예정이다.

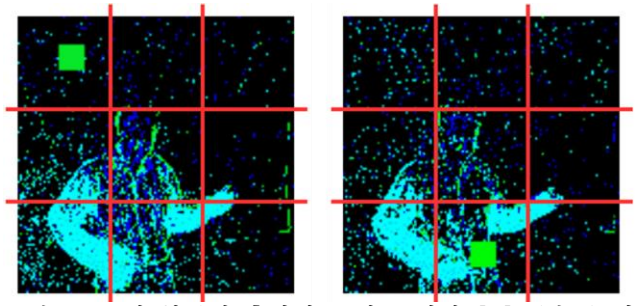

그림 1: 특정 블록에 추가된 극성 트리거 [1]. (왼쪽) 가장 이벤트가 드문 블록에 추가된 극성 트리거. (오른쪽) 가장 이벤트가 잦은 블록에 추가된 극성 트리거.

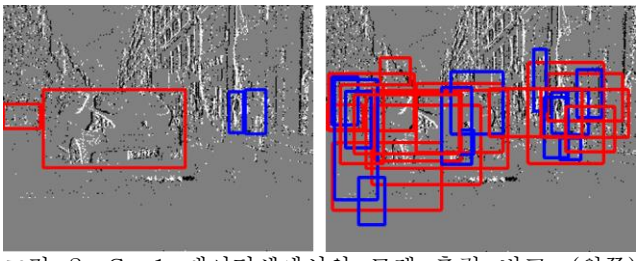

그림 2: Gen1 데이터셋에서의 모델 출력 비교. (왼쪽) [8] 정상 입력에 대한 객체 후보 출력. (오른쪽) 제안된 트리거가 삽입된 입력에 대한 가짜 객체 후보 출력.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학석사연계 ICT 핵심인재양성사업의 연구결과로 수행되었음. (IITP-2025-RS-2024-00437024) 교신저자: 이은규 (eklee@inu.ac.kr)

참 고 문 헌

[1] Abad, G, et al. Sneaky spikes: Uncovering stealthy backdoor attacks in spiking neural networks with neuromorphic data. arXiv preprint arXiv:2302.06279, 2023. https://github.com/GorkaAbad/Sneaky-Spikes/tree/main#

[2] DE TOURNEMIRE, Pierre, et al. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020.

[3] Jin, L, et al. Data Poisoning-based Backdoor Attack Framework against Supervised Learning Rules of Spiking Neural Networks. arXiv preprint arXiv:2409.15670, 2024.

[4] Kim, S, et al. Spiking-yolo: spiking neural network for energy-efficient object detection. AAAI, 2020. p. 11270-11277.

[5] Luo, X, et al. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. ECCV, Cham: Springer Nature Switzerland, 2024. p. 253-272.

[6] Su, Q, et al. Deep directly-trained spiking neural networks for object detection. IEEE/CVF ICCV. 2023. p. 6555-6565.

[7] Xiao, Y, et al. Sponge Backdoor Attack: Increasing the Latency of Object Detection Exploiting Non-Maximum Suppression. IEEE IJCNN. 2024. p. 1-8.

[8] "Prophesee Gen1 Automotive Detection Dataset", https://www.prophesee.ai/2020/01/24/prophesee-gen1-automotive-detection-dataset/