국방 분야에서 이미지 기반 딥러닝 모델에 대한 적대적 샘플에 관한 고찰

정찬호, 권 현*
육군사관학교 사이버전 전공, *육군사관학교 전자공학과
cksghsla67@gmail.com, *hkwon.cs@gmail.com

A Study on Adversarial Examples for Image-Based Deep Learning Model in the Military

Chan Ho Jeong, Hyun Kwon*
Korea Millitary Academy., *Dept. of Electrical Engineering, Korea Military Academy.

요 약

딥뉴럴네트워크를 이용하여 이미지 인식 분야에서 좋은 성능을 보여주고 있다. 이미지 분야에 딥뉴럴네트워크를 이용하여 국방 분야에서 사물을 인식하거나 탐지하는 분야에 접목하여 사용되고 있다. 하지만 딥뉴럴네트워크에 대한 취약점 중에 적대적 샘플에 대한 위험성이 존재한다. 적대적 샘플은 원본 이미지에 최소한의 왜곡을 추가하여 사람에 의해서 정상적인 이미지로 식별되지만 모델에 의해서 잘못 오인식되는 샘플을 의미한다. 이러한 적대적 샘플이 국방분야에 적용한 연구가 미흡한 실정이다. 본 논문에서 이미지 분야에서 적대적 샘플에 대한 발생할 수 있는 상황, 실험적 결과 및 국방 분야에 대한 고찰에 대하여 분석하였다.

Ⅰ. 서 론

컴퓨팅 기술이 발전되고 클라우드 환경을 통한 빅데이터 수집이 가능하게 되면서 인공지능 기술이 발전되고 있다. 더불어서 4차 산업혁명과 맞춰서 인공지능 기술은 시너지 효과를 내고 있으며 국가 차원에서 산업계, 교육계 등에서 인공지능 기술을 접목하여 활용하는 서비스와 연구가 활발하다. 그러한 인공지능 기술 중에 딥뉴럴네트워크[1]가 있다. 딥뉴럴네트워크는 이미지 인식, 텍스트 인식, 음성 인식 등에 좋은 성능을 제공하고 있다. 특히, 이미지 인식 분야에 있어서 사람에 버금가거나 능가하는 성능을 제공하고 있어 산업계와 교육계에서 다양한 연구와 서비스를 제공하고 있다. 국방 분야에서 이러한 이미지 기반에 딥뉴럴네트워크를 이용하여 과학화 감시장비 체계에 접목하는 연구를 진행하고 있다.

하지만 이러한 딥뉴럴네트워크는 적대적 샘플 공격[2]에 대하여 취약점이 존재한다. 적대적 샘플 공격은 원본 이미지에 약간의 노이즈를 추가하여 사람에 보기에는 정상적인 이미지이지만 모델에 의해서 잘못 오인식되는 샘플을 의미한다. 이러한 적대적 샘플을 국방분야에 접목할 경우, 오인식을 일으킬 가능성이 높다. 따라서 이에 대한 방어 대책 등에 대한 연구가 필요하다.

본 논문에서는 국방 분야[3]에서 사용될 수 있는 딥뉴럴네트워크에 대한 적대적 샘플 공격과 실험적인 가능성 그리고 향후 적용하는 방안에 대하여 연구하였다.

Ⅱ. 이미지 기반 딥러닝 모델에서의 적대적 샘플 공격

이미지 기반의 적대적 샘플을 생성하기 위해서 타겟 모델은 VGG19 모델[4]을 대상으로 실험을 하였다. 데이터셋은 ImageNet [5]기반으로 총 1000가지의 클래스별로 구성되어있다. ImageNet을 통해서 학습된 모델을 대상으로 국방에서 활용되는 전투 사진에 대하여 적대적 샘플 생성방법은 shadow adversarial example 방법[6]을 적용하였다. 이 방법은 입력 이미지에 대한 모델의 confidence 값을 확인 후에 잘못 오인식되는 클래스의 confidence 값이 가장 높을 때까지 노이즈를 추가하여 생성하는 방법이다. 특히, 이 방법은 사람이 식별하기 어려운 이미지의 boundary 영역에만 노이즈를 추가하여 좀 더 이미지 왜곡을 사람이 보기에 적게 만든 특징이 있다. 이 방법에 대한 도식은 그림 1과 같다.

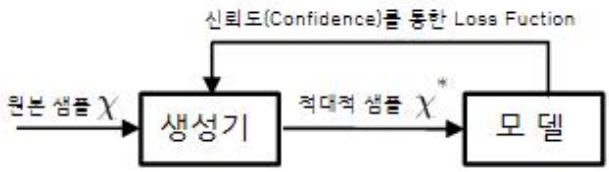

그림 1. 적대적 샘플 생성

실험 결과 측면에서, 이 방법을 적용하여 국방 분야에서 사용될 수 있는 전투기 사진에 대하여 아래와 같이 가능성에 대한 실험을 하였다.

그림 2. 전투기 사진

```
< 가장 높은 확률을 가지는 클래스들 >
인덱스: 895 / 클래스명: warplane, military plane / 확률: 75.3515%
인덱스: 908 / 클래스명: wing / 확률: 20.0965%
인덱스: 134 / 클래스명: crane / 확률: 0.3937%
인덱스: 701 / 클래스명: parachute, chute / 확률: 0.3911%
인덱스: 99 / 클래스명: goose / 확률: 0.199%
```

그림 3. 그림 2의 원본 전투기 사진에 대한 분류 결과

그림 2는 원본 샘플로서 전투기 사진 예시를 보여주며 그림 3은 그림 2에 대한 각 클래스별로 가장 높은 확률값을 갖는 클래스별로 보여준다. 그림 3에서 보면 military plane으로 군사 항공기로 제대로 분류된 것을 볼 수가 있다. 인덱스는 총 1000가지로 해당되는 클래스의 번호를 의미한다. 이 전투기 사진에 대한 적대적 샘플을 생성하게 되면 그림 4과 그림 5와 같이 결과가 나온다.

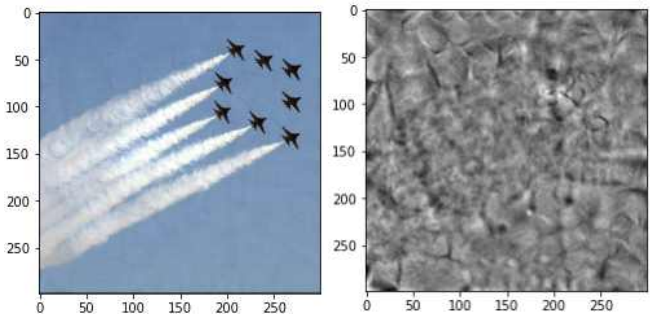

그림 4. 적대적 샘플로 생성한 전투기 사진

```
< 가장 높은 확률을 가지는 클래스들 >
인덱스: 18 / 클래스명: magpie / 확률: 99.9724%
인덱스: 17 / 클래스명: jay / 확률: 0.0073%
인덱스: 80 / 클래스명: black grouse / 확률: 0.006%
인덱스: 143 / 클래스명: oystercatcher, oyster catcher / 확률: 0.0026%
인덱스: 981 / 클래스명: ballplayer, baseball player / 확률: 0.0023%
```

그림 5. 그림 4의 적대적 샘플 사진에 대한 분류 결과

그림 4는 적대적 샘플에서 생성한 전투기 사진과 이미지 상에 왜곡이 들어간 부분을 보여준다. 그림에서 보면 특정 boundary 영역에서 왜곡이 들어간 것을 볼 수가 있다. 또한 그림 4에서의 전투기 사진은 사람이 보기에 그림 2에서의 원본 전투기 사진과 거의 유사한 것을 볼 수가 있다. 하지만 그림 5와 같이 모델에 의해서 이 적대적 샘플로 생성된 전투기 사진은 까치(mapie)로 잘못 오인식된 것을 볼 수가 있다. 이와 같이 적대적 샘플을 국방 관련 이미지 데이터에 적용하였을 때, 모델에 의해서 잘못 오인식된 것을 볼 수가 있다.

Ⅲ. 국방 분야에서의 적대적 샘플을 적용한 고찰

국방 분야에서 이러한 적대적 샘플을 적용한 방법들이 이용될 수가 있다. 가장 대표적으로 활용될 수 있는 영역은 기만과 위장 영역이다. 먼저 적군을 기만하는 예시로 발사 후 표적을 고착하는 LOAL 자동인식 소프트웨어들을 교란시킬 수 있다. AI를 이용한 무인화된 자율적 표적식별 시스템[7]을 교란시킴으로써 우리 군의 주요 지휘부가 집중적으로 공격당하는 것을 방어하는 효과를 창출해낼 수 있다. 또한 무선통신 영역에서는 적절한 잡음을 섞어 적대적 사례를 발생시킴으로써 적군의 도청을 방지하는 등의 효과를 기대할 수 있다. 위장에 활용되는 예시로는, 디지털 복장이나 전차와 같은 표면에 딥러닝 모델에 의해서 잘못 오인식 되는 노이즈를 추가하여 생성할 수가 있다. 특히 adversarial patch와 같은 다소 왜곡이 강하더라도 black box 형태의 딥러닝 모델에서도 공격이 가능한 특징이 있다. 방어적 측면에서 과학화 감시장비[8]에서 딥뉴럴네트워크 기술이 접목될 경우, 적대적 샘플 방어방법에 대한 기법이 반영되어야 한다. 대치하고 있는 적군이 특정 복장을 통하여 딥러닝 모델을 오인식하여 침투를 허용하는 등의 문제가 야기되기 때문에 이에 대한 연구가 병행되어야 한다. 실제로 미국의 경우, DARPA에서 적대적 샘플 공격에 취약점을 보안하기 위한 국방 딥러닝 모델에 대한 연구가 진행되고 있는 상황이다. 현재 국방에서는 딥러닝 모델을 적용하기 위한 서비스 측면에서 연구가 많이 진행되고 있지만 이러한 보안 분야에 대한 고려도 필요하다.

Ⅳ. 결론

본 논문에서 국방분야에 적용 될 수 있는 이미지 기반의 딥뉴럴네트워크에 대한 적대적 샘플 공격과 실험적 분석 그리고 국방 분야에 적용에 대한 고찰을 다루었다. 향후 이러한 국방 분야에 딥러닝 기술을 이미지 분야에 접목할 경우, 방어체계에 대한 고려가 필요하다. 향후 연구는 이러한 국방 분야에 적대적 샘플을 방어할 수 있는 분야에 대하여 연구를 진행할 예정이다.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2021R1I1A1A01040308)

참 고 문 헌

[1] Liu, Weibo, et al. "A survey of deep neural network architectures and their applications." Neurocomputing 234 (2017): 11-26.

[2] Kwon, Hyun, et al. "Advanced ensemble adversarial example on unknown deep neural network classifiers." IEICE RANSACTIONS on Information and Systems 101.10 (2018): 2485-2500.

[3] Kwon, Hyun, et al. "Friend-safe evasion attack: An adversarial example that is correctly recognized by a friendly classifier." computers & security 78 (2018): 380-397.

[4] Jaworek-Korjakowska, Joanna, Pawel Kleczek, and Marek Gorgon. "Melanoma thickness prediction based on convolutional neural network with VGG-19 model transfer learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[5] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[6] Ghiasi, Amin, Ali Shafahi, and Tom Goldstein. "Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates." arXiv preprint arXiv:2003.08937 (2020).

[7] 신종호, et al. "무인수상정의 자율운항을 위한 시스템 식별 및 제어." 대한기계학회 춘추학술대회 (2015): 41-42.

[8] 최치원, 송태식, and 엄정호. "과학화 장비를 활용한 무인 보안시스템 운영 방안에 관한 연구." 보안공학연구논문지 9.3 (2012): 209-218.