JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 10, pp. 67-76, October 2023
https://doi.org/10.9708/jksci.2023.28.10.067

Detecting Adversarial Examples Using Edge-based Classification

Jaesung Shim*, Kyuri Jo*

*Student, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea
*Professor, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea

[Abstract]

Although deep learning models are making innovative achievements in the field of computer vision, the problem of vulnerability to adversarial examples continues to be raised. Adversarial examples are attack methods that inject fine noise into images to induce misclassification, which can pose a serious threat to the application of deep learning models in the real world. In this paper, we propose a model that detects adversarial examples using differences in predictive values between edge-learned classification models and underlying classification models. The simple process of extracting the edges of the objects and reflecting them in learning can increase the robustness of the classification model, and economical and efficient detection is possible by detecting adversarial examples through differences in predictions between models. In our experiments, the general model showed accuracy of {49.9%, 29.84%, 18.46%, 4.95%, 3.36%} for adversarial examples (eps={0.02, 0.05, 0.1, 0.2, 0.3}), whereas the Canny edge model showed accuracy of {82.58%, 65.96%, 46.71%, 24.94%, 13.41%} and other edge models showed a similar level of accuracy also, indicating that the edge model was more robust against adversarial examples. In addition, adversarial example detection using differences in predictions between models revealed detection rates of {85.47%, 84.64%, 91.44%, 95.47%, and 87.61%} for each epsilon-specific adversarial example. It is expected that this study will contribute to improving the reliability of deep learning models in related research and application industries such as medical, autonomous driving, security, and national defense.

▸ **Key words**: Deep Learning, Computer Vision, Convolutional Neural Network, Edge-based Classification, Adversarial Example Detection

• First Author: Jaesung Shim, Corresponding Author: Kyuri Jo
*Jaesung Shim (climbsky@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
*Kyuri Jo (kyurijo@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
• Received: 2023. 09. 15, Revised: 2023. 10. 19, Accepted: 2023. 10. 19.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

딥러닝 모델이 컴퓨터 비전 분야에서 혁신적인 성과를 이루어내고 있으나, 적대적 예제에 취약하다는 문제가 지속적으로 제기되고 있다. 적대적 예제는 이미지에 미세한 노이즈를 주입하여 오분류를 유도하는 공격 방법으로서, 현실 세계에서의 딥러닝 모델 적용에 심각한 위협이 될 수 있다. 본 논문에서는 객체의 엣지를 강조하여 학습된 분류 모델과 기본 분류 모델 간 예측 값의 차이를 이용하여 적대적 예제를 탐지하는 모델을 제안한다. 객체의 엣지를 추출하여 학습에 반영하는 과정만으로 분류 모델의 강건성을 높일 수 있으며, 모델 간 예측값의 차이를 통하여 적대적 예제를 탐지하기 때문에 경제적이면서 효율적인 탐지가 가능하다. 실험 결과, 적대적 예제 (eps={0.02, 0.05, 0.1, 0.2, 0.3})에 대한 일반 모델의 분류 정확도는 {49.9%, 29.84%, 18.46%, 4.95%, 3.36%}를 보인 반면, Canny 엣지 모델은 {82.58%, 65.96%, 46.71%, 24.94%, 13.41%}의 정확도를 보였고 다른 엣지 모델들도 이와 비슷한 수준의 정확도를 보여, 엣지 모델이 적대적 예제에 더 강건함을 확인할 수 있었다. 또한 모델 간 예측값의 차이를 이용한 적대적 예제 탐지 결과, 각 epsilon별 적대적 예제에 대하여 {85.47%, 84.64%, 91.44%, 95.47%, 87.61%}의 탐지율을 확인할 수 있었다. 본 연구가 관련 연구 분야 및 의료, 자율주행, 보안, 국방 등의 응용 산업 분야에서 딥러닝 모델의 신뢰성 제고에 기여할 것으로 기대한다.

▸ **주제어**: 딥러닝, 컴퓨터 비전, 합성곱 신경망, 엣지 기반 분류, 적대적 예제 탐지

I. Introduction

딥러닝 모델은 이미지 분류 및 객체 탐지 등 다양한 컴퓨터 비전 분야에서 뛰어난 성능을 보여주고 있다. 그러나 최근의 연구들은 이러한 딥러닝 모델이 적대적 예제에 취약하다는 문제를 제기하고 있다. 적대적 예제란 변형이 없는 원본 이미지(이하 클린 이미지)에 의도적으로 미세한 노이즈(perturbation)를 주입함으로써 분류 모델의 잘못된 예측을 유도하도록 생성된 이미지 데이터를 말한다[1].

적대적 예제는 실제 환경에서의 딥러닝 모델 적용에 중대한 위협이 될 수 있다. 예를 들어, 적대적 예제를 이용한 공격은 자율주행 자동차의 잘못된 판단을 유도하거나 보안 검사를 피해 침입하는 등의 사회적, 경제적인 문제를 야기할 수 있다[2, 3, 4].

이와 관련하여, 딥러닝 모델의 취약성을 분석하고, 새로운 방식의 적대적 예제를 생성하고, 이에 대한 대응 방안을 제시한 많은 연구들이 있다[5]. 다양한 대응 방안 중 적대적 예제를 생성하여 모델이 학습하도록 하는 적대적 훈련은 적대적 예제에 효과적이라고 알려져 있지만[1, 6], 적대적 예제의 생성 및 학습은 복잡하고 계산적으로 비용이 많이 드는 작업이며 모델의 정확도가 저하되는 단점이 있다[7].

Borji는 원본 이미지에 객체의 엣지를 더하여 분류 모델을 학습시킨다면 적대적 예제에 더 강건해 질 수 있다는 연구 결과를 제시하였다[8]. 여러 종류의 데이터를 대상으로 다양한 엣지 결합 방법을 적용하였으며, 그림 1과 같은 평균 결과를 도출하였다.

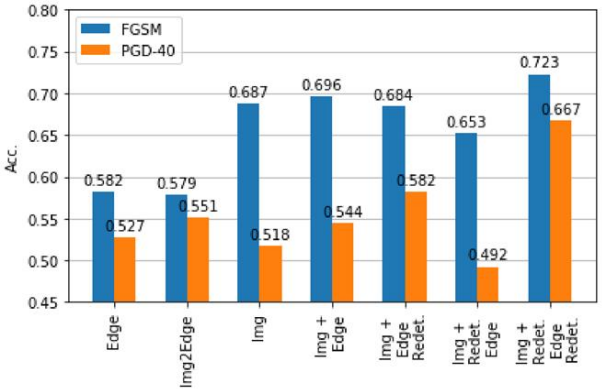

Fig. 1. Average Results of EAT(Edge-guided Adversarial Training) Defense[8]

그리고, Xu et al.은 Feature Squeezing 모델과의 예측값 비교를 통한 적대적 예제 탐지 방법을 제안하였는데[9], 다양한 알고리즘이 적용된 적대적 예제들에 대한 탐지율은 그림 2와 같다.

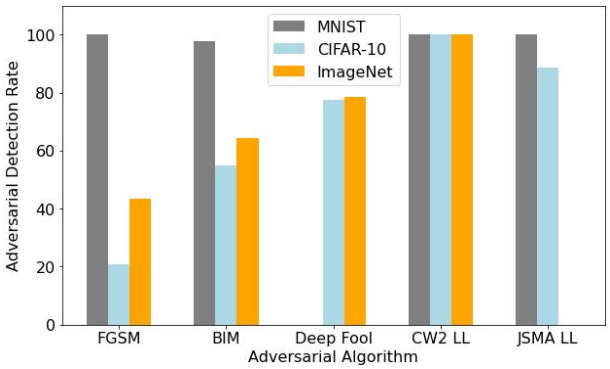

Fig. 2. Detection Rates of Feature Squeezing Model

본 논문에서는 위의 선행 연구들[8, 9]을 기반으로 객체의 엣지를 강조하여 학습한 분류 모델과 기본 분류 모델 간의 분류 예측값의 차이를 이용하여 적대적 예제를 간단하면서도 효과적으로 탐지하는 방법을 제안한다.

본 연구를 통해, 적대적 예제에 대한 딥러닝 모델의 강건성을 향상시키고 보안적인 측면에서의 신뢰성을 제고할 수 있으며, 이를 기반으로 실제 응용 환경에서 딥러닝 분류 모델을 보다 안전하게 활용할 수 있을 것이다.

본 논문의 2장에서는 적대적 예제 생성 및 방어 기법과 엣지 탐지 기법들을 소개하고, 3장에서는 제안하고자 하는 적대적 예제 탐지 모델을 제시한 후, 4장에서 실험 과정 및 결과를 설명하고, 5장에서 결론을 제시하였다.

II. Preliminaries

1. Related works

1.1 Adversarial Example

컴퓨터가 처리하는 이미지는 일반적으로 한 픽셀당 빛의 삼원색인 빨강(R), 초록(G), 파랑(B)에 대한 값을 가지고 있으며, RGB 각각은 8bits의 정보, 즉 256개(0~255)의 정보로 표현된다. 사람은 해당 정보의 미세한 변화를 구분할 수 없지만, 딥러닝 모델에서는 복잡한 계층으로 이루어진 네트워크를 통과하면서 가중치가 곱해지고 차원이 증가하기 때문에 처음의 미세한 변화가 증폭되어 결과적으로 이미지에 대한 오분류를 일으키게 된다[1]. 적대적 예제는 이렇게 딥러닝 모델이 오분류를 일으키도록 클린 이미지에 아주 작은 노이즈를 추가하여 생성된 이미지이다.

적대적 예제를 생성하기 위해 사용되는 주요 기법들은 다음과 같다.

- Fast Gradient Sign Method (FGSM) : Ian Goodfellow 등이 2014년에 제안한 대표적인 적대적 예제 생성 기법으로, 모델의 손실 함수를 미분하여 구해진 기울기의 부호를 이용하여 적대적 예제를 생성한다[1]. 클린 이미지에 사람이 인지하기 어려운 작은 노이즈를 추가하는 것만으로 모델의 오분류를 유도하는 간단하고 효과적인 기법이다.

- Projected Gradient Descent (PGD) : Aleksander Madry 등이 2017년에 제안한 FGSM의 개선 버전으로서, 단일 스텝으로 적대적 예제를 생성하는 FGSM과 달리, PGD는 일정한 횟수(스텝)를 반복하여 적대적 예제를 생성하는 기법이다[6]. 반복횟수 및 학습율을 통해서 클린 이미지를 조금씩 업데이트하여 더 강력한 적대적 예제를 생성할 수 있으나, 많은 계산 리소스를 필요로 한다.

- DeepFool : Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard가 2016년에 제안한 적대적 예제 생성 기법으로서, L2 Norm으로 계산된 가장 가까운 경계 평면으로 클린 이미지를 이동시켜 클래스를 오분류하도록 반복적으로 업데이트하는 방식이다[10]. 최적의 적대적 예제를 찾아낼 수 있다는 장점이 있으나, 많은 계산 리소스를 필요로 한다.

- Carlini & Wagner (C&W) : Nicholas Carlini와 David Wagner가 2017년에 제안한 강력한 적대적 예제 생성 기법으로서, L2, L0, L∞ 기반 Distance Metric을 각각 사용한 세가지 공격방식을 제시하였다[11]. 클린 이미지와 적대적 예제의 차이를 최소화하면서 비선형적인 제약조건을 만족시키는 원리로서, FGSM이나 PGD 기법과 달리 L2, L0, L∞ 모두를 사용한 공격이 가능하며 공격 성공률도 매우 우수한 기법으로 심층 신경망(deep neural network, 이하 DNN) 분류 모델의 강건성 평가 방법으로 많이 사용되어져 왔다.

1.2 Adversarial Defense

적대적 예제에 대한 다양한 방어 기법들 가운데 본 연구와 관련된 방어 기법으로, 적대적 예제에 대한 DNN 분류 모델의 강건성을 직접 향상시키는 적대적 훈련 방법과 이미지 분류 전에 적대적 예제 여부를 탐지하는 적대적 예제 탐지 방법이 있다.

- 적대적 훈련(Adversarial Training) : 딥러닝 모델이 적대적 예제를 직접 학습함으로써 더 강건하고 안정적으로 동작하도록 하는 방법이다[1, 6]. 적대적 예제 생성 네트워크(Generative Adversarial Networks)를 통하여 생성된 적대적 예제를 딥러닝 모델이 학습하도록 하여 결과적으로 적대적 예제에 강건하게 동작하도록 한다.

적대적 훈련은 알려진 적대적 공격에 대하여 효과적인

방어 수단이지만, 적대적 예제 생성 및 학습을 위해 많은 리소스가 필요하며, 클린 이미지에 대한 정확도가 낮아지는 문제, 그리고 새로운 기법으로 생성된 적대적 예제에 대해서는 방어가 어렵다는 단점이 있다.

- 적대적 예제 탐지(Adersarial Detection) : 딥러닝 모델이 이미지에 대한 분류 예측 결과를 판단하기 전에, 정상 이미지인지 아니면 적대적 예제인지를 먼저 판별하도록 하는 방법이다[9, 12, 13]. 정상 이미지라면 분류 예측 결과를, 적대적 예제라면 적대적 예제임을 출력한다.

적대적 예제 탐지 기법 중 '예측 불일치(prediction inconsistency)' 기법은 하나의 적대적 예제가 모든 딥러닝 모델을 속일 수 없다는 가정을 전제로 하여, 여러 모델에게 입력 이미지를 예측하도록 한 후 그 예측 결과 간의 일치 여부를 확인하는 방법[9, 12]으로서, 성능 면에서 효과적이며 비교적 리소스가 적게 드는 장점이 있다. 본 논문에서 제안하는 탐지 모델이 이 기법에 속한다.

1.3 Edge Detection

사람은 이미지 내의 객체를 인식할 때 객체의 형태에 더 의존하는 반면, 딥러닝 모델은 객체의 텍스처에 더 편향된 인식을 한다. 이것은 엣지를 구성하는 픽셀보다 텍스처를 구성하는 픽셀의 수가 많기 때문이며, 결과적으로 적대적 공격의 기회를 제공하게 된다. 딥러닝 모델에게 객체의 엣지 부분을 더 강조하여 인식하도록 학습한다면 적대적 공격에 강건함을 유지할 수 있을 것이다[8].

이미지에서 객체의 엣지 탐지(Edge Detection)는 일반적으로 픽셀 값이 급격하게 변하는 지점을 객체의 경계로 인식하여 표현한다.

- 로버츠 크로스 엣지 탐지(Roberts Cross Edge Detection) : 1963년에 L. Roberts가 제안한 방법으로, 2x2 크기의 작은 커널을 사용하여 이미지의 대각선 방향의 변화율(기울기)을 계산하고, 이를 통해 엣지를 검출한다[14]. 연산이 간단하고 빠르며, 특히 대각선 방향의 엣지를 잘 검출하는 특징이 있다.

- 소벨 엣지 탐지(Sobel Edge Detection) : 1968년에 I. Sobel과 G. Feldman이 제안한 방법으로, 3x3 소벨 필터를 이용하여 이미지의 픽셀 값들의 변화율(기울기)을 계산하여 엣지를 검출한다[15]. 이미지의 x축과 y축 방향으로 소벨 필터를 적용하여 얻어진 기울기로부터 강도와 방향을 구한 후 이를 이용하여 엣지를 판단한다.

- 캐니 엣지 탐지(Canny Edge Detection) : 1986년에 J. Canny가 제안한 방법[16]으로, 가장 유명하고 많이 사용된다. 노이즈 감소, 그래디언트 계산, 비최대 억제, 이중 임계값, 히스테리시스 임계값 처리 등의 다단계 프로세스를 거쳐 엣지를 검출하는데, 이를 통해 정확하고 선명한 엣지를 찾아낼 수 있다.

- 샤르 엣지 탐지(Scharr Edge Detection) : 2000년에 H. Scharr가 제안한 방법으로, 소벨 엣지 탐지 기법과 동일한 방식이지만, 필터 값을 다르게 적용하여 소벨 탐지 기법보다 정밀한 결과를 얻을 수 있다[17].

III. The Proposed Scheme

제안하는 탐지 모델은 일반적인 분류 모델과 이미지 내 객체의 엣지를 강조하여 학습한 분류 모델(이하 엣지 모델) 간의 예측 결과를 비교하여 적대적 예제 여부를 판별하며, 이를 위하여 크게 학습 및 예측, L1 Distance 계산, 적대적 예제 판별의 3단계로 진행된다(그림 3).

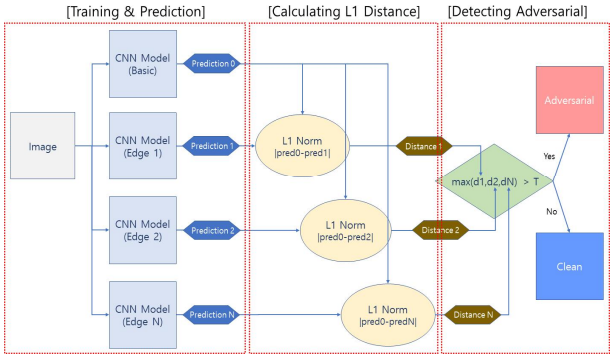

Fig. 3. Adversarial Example Detection Process

1. Training & Prediction

이미지 분류에 주로 활용되는 합성곱 신경망(convolutional neural network, CNN) 모델을 기본 모델로 선정하여 클린 이미지를 대상으로 학습을 수행한다. 기본적인 분류 정확도가 적대적 예제 탐지 성능에 직접적인 영향을 주기 때문에 최대한 높은 정확도를 획득할 수 있도록 학습을 수행하는 것이 중요하다.

엣지 모델 학습은 이미지로부터 추출한 엣지를 다시 이미지에 결합한 뒤 분류 모델에 입력하여 학습케 하는 전처리(pre-processing) 방식으로 수행한다. 먼저 엣지 추출 기법을 사용하여 이미지로부터 엣지 이미지를 추출한 후, RGB 3채널로 구성된 원본 이미지 데이터에 4번째 채널로 결합시킨다(그림 4). 이렇게 객체의 엣지가 강조된 이미지를 분류 모델에 입력하여 반복적으로 학습을 진행한다.

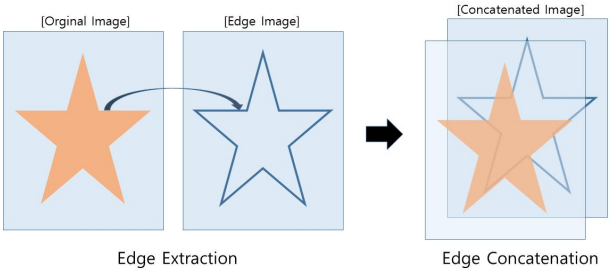

Fig. 4. Edge Extraction & Concatenation

본 탐지 모델에서는 복수 개의 엣지 모델을 사용하므로, 서로 다른 여러 엣지 추출 기법을 사용하여 추가적인 엣지 모델을 학습시킨다.

입력된 이미지에 대한 모델 간 L1 Distance를 계산하기 위하여, 기본 모델과 엣지 모델들의 클래스 별 정답 확률을 소프트맥스 함수를 사용하여 산출한다. 일반적인 분류 모델에서 정답 클래스를 최종 예측하기 위한 바로 전 단계의 값이다. 각 클래스별 확률 값은 0~1 사이의 값이며 모두 더하면 1이 된다.

2. Calculating L1 Distance

L1 Norm을 적용하여 기본 모델과 엣지 모델의 예측값(클래스 별 확률 값) 간의 거리를 계산한다.

$$\text{L1 Distance} = \left(\sum_{i=1}^{n} \left|(\text{Prediction}(x)-\text{Prediction}^{\text{edged}}(x))_i\right|\right)$$

L1 Norm은 벡터 요소들의 절대값의 합으로 계산되는데, 본 실험에서는 위의 식처럼 기본 모델의 각 클래스별 확률 값들과 엣지 모델의 각 클래스별 확률 값들의 차를 구한 뒤 이를 벡터 요소들로 하여 절대값의 합을 계산한다. 참고로, 각 클래스 별 확률의 최대값은 1, 최소값은 0, 확률의 총 합계는 1이기 때문에, 결과적으로 L1 Distance는 0~2 사이의 값을 갖게 된다.

3. Detecting Adversarial

입력 이미지의 적대적 예제 여부를 판단하기 위한 값으로 기본 모델과 복수개의 엣지 모델 간 L1 Distance 중에서 가장 큰 값을 선택한다. 이것은 적대적 예제가 모든 엣지 모델을 속이기는 어려울 것이라는 가정하에, 하나의 L1 Distance에서라도 적대적 예제의 가능성을 탐지한다면 그에 근거하여 적대적 예제로 판단한다는 의미이다. 그러나, 반대로 클린 이미지에 대하여 하나의 L1 Distance에서 오판을 한다면 그 역시 적대적 예제로 판단하여, 결과적으로 오탐을 하게 될 위험성도 있다. 그러므로 본 탐지 모델에서는 기본 모델과 엣지 모델의 기본적인 분류 성능이 매우 중요하다.

Max L1 Distance가 임계치(Threshold)를 초과하면 적대적 예제로 분류한다. 여기서 임계치란 적대적 예제를 탐지하기 위하여 입력 이미지의 Max L1 Distance가 적대적 예제의 범위에 속하는가를 판단할 임의의 기준값을 의미한다. L1 Distance에 의해 적대적 예제를 판단하기 때문에, 임계치를 몇으로 설정하느냐에 따라서 적대적 예제 탐지율이 높아지기도 하고 낮아지기도 한다. 임계치가 낮으면 적대적 예제 탐지율(진양성, True Positive)은 높아지지만 클린 이미지를 적대적 예제로 판단하는 오탐(위양성, False Positive)이 증가하게 된다. 반대로 임계치가 높으면 오탐이 줄어드는 대신 적대적 예제를 클린 이미지로 판단(위음성, False Negative)하게 되어 적대적 예제 탐지율이 낮아지게 된다(Trade-off).

IV. Experiments and Results

적대적 예제 탐지 모델의 구현 및 검증은 데이터 셋 구축, 기본 모델 및 엣지 모델 학습, 적대적 예제 생성 및 엣지 모델 강건성 확인, L1 Distance 계산 및 임계치 도출, 적대적 예제 탐지의 순으로 진행하였다.

1. Preparing Data Set

30종의 동물 이미지[18]중에서, 기본 분류 모델로 이용할 EfficientNet-B1 모델의 입력 이미지 사이즈(240x240)와 실험 장비의 성능을 고려하여, 가로 또는 세로의 크기가 200pixel 이상이면서 용량이 200kb이하인 이미지들로 20,210개를 선별하였다(그림 5).

Fig. 5. Examples of Image Data Set

정확하고 객관적인 실험이 될 수 있도록, 훈련(Training), 검증(Validation), 평가(Test) 데이터를 7:2:1의 비율로 분류한 후, 반복되는 훈련 및 평가 과정에서 섞이지 않도록 하였다.

2. Model Training & Evaluation

2.1 Model Training

기본 모델과 엣지 모델의 분류 예측 결과값의 비교를 통하여 적대적 예제를 탐지하고자 하는 본 연구에서, 기

본 모델의 분류 예측 정확도가 높지 않다면 그로부터 기인된 오탐이 발생하게 되므로, 정확한 실험을 위해서는 분류 예측 정확도를 높이는 작업은 매우 중요하다.

본 연구에서는 실험 장비의 성능을 고려하여 정확도가 높으면서도 파라미터 수가 적은 EfficientNet-B1 모델을 기본 모델로 선정하였으며, 준비된 30종의 동물 이미지 데이터 셋을 사용하여 전이학습을 수행하였다. 학습을 마친 후 평가 데이터를 통한 기본 모델의 분류 예측 정확도를 테스트한 결과 약 92%(92.97%)의 정확도를 확인할 수 있었다.

이미지 내 객체의 엣지를 탐지하기 위하여 Canny, Sobel, Scharr 엣지 탐지 방법을 적용하였다. 실제 구현은 OpenCV 패키지에서 제공하는 엣지 함수를 사용하였으며, 적대적 예제에 대한 분류 정확도를 높이기 위해서 함수의 입력 파라미터를 적절히 조정하였다(그림 6). 특히 강한 노이즈가 주입된 적대적 예제의 경우는 노이즈도 엣지로 탐지하기 때문에 강한 엣지만을 탐지하도록 파라미터를 상향 조정하여 오탐을 최대한 방지하였다.

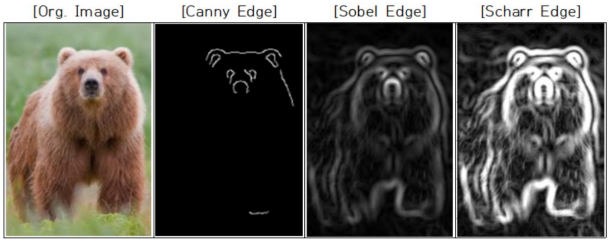

Fig. 6. Edge Images (Parameters have been adjusted to detect only strong edges.)

30종의 동물 이미지들에 대하여, 추출된 엣지 이미지를 원본 이미지에 4번째 채널로 추가한 후 분류 모델에 입력하는 방식으로 엣지 학습을 진행하였다. 학습을 마친 후 평가 데이터로 테스트한 결과, 표 1과 같이 기본 모델과 유사한 92%~93%의 분류 정확도를 확인할 수 있었다.

Table 1. Accuracy of Edge Models

Model	Acc.
Canny Model	93.22%
Sobel Model	92.28%
Scharr Model	92.33%

2.2 Adversarial Examples & Evaluation

적대적 예제를 생성하기 위하여 평가 데이터에 FGSM 기법을 적용하였으며, 실제 구현은 Torchattacks 패키지에서 제공하는 함수를 사용하였다. 적대적 예제의 노이즈 강도를 결정하는 epsilon은 [0.02, 0.05, 0.1, 0.2, 0.3]으로 설정하였다(그림 7).

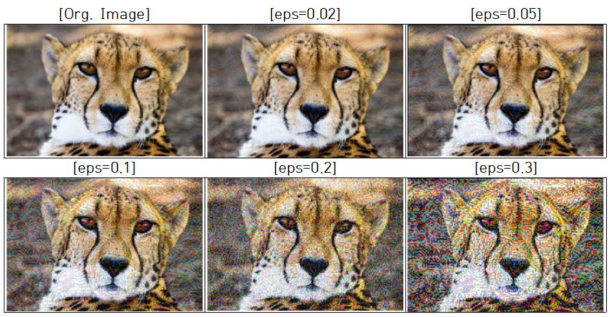

Fig. 7. Adversarial Images by Epsilon

적대적 예제 탐지를 위한 전제 요건인 엣지 모델의 강건성을 확인하기 위하여 적대적 예제에 대한 각 엣지 모델 별 분류 예측 정확도를 평가하였다.

Table 2. Accuracy for Adversarial Examples

eps	Basic Model	Canny Model	Sobel Model	Scharr Model
0.02	49.93%	82.58%	80.46%	83.62%
0.05	29.84%	65.96%	64.57%	69.27%
0.1	18.46%	46.71%	46.07%	50.96%
0.2	4.95%	24.94%	19.25%	28.90%
0.3	3.36%	13.41%	8.26%	13.81%

표 2를 보면, 적대적 예제에서 epsilon을 0.02로 설정하였을 때, 기본 모델의 분류 예측 정확도는 49.93%로 낮아진 반면, 엣지 모델들은 80% 이상의 높은 정확도를 유지하였다. epsilon을 0.05로 설정하였을 때, 기본 모델의 분류 예측 정확도는 29.84%로 매우 낮아졌으나, 엣지 모델들은 60% 이상의 정확도를 유지하였다.

결과적으로, 객체의 엣지를 강조하여 학습한 모델들이 적대적 예제에 대하여 매우 강건함을 확인할 수 있다.

3. L1 Distance

3.1 Calculating L1 Distance

앞서 기술한 것처럼 입력 이미지에 대한 기본 모델과 엣지 모델 간 예측 값의 차이(거리)는 두 모델 간의 각 클래스 별 확률 값들의 차이에 L1 Norm을 적용하여 계산하였다.

표 3은 클린 이미지와 그 이미지를 원본으로 하여 생성한 적대적 예제들에 대한 기본 모델과 각 엣지 모델 간 L1 Distance를 계산한 결과이다. 클린 이미지(epsilon=0.00)에 대하여 기본 모델과 엣지 모델들이 정

답 클래스인 2를 예측했으며, 그 결과 L1 Distance들이 0.01 미만의 값임을 알 수 있다. epsilon을 0.02, 0.05, 0.1으로 설정한 적대적 예제들에 대하여, 기본 모델은 오답 클래스를 예측하였으나 엣지 모델들은 정답 클래스를 예측하였는데, 그 결과 L1 Distance들이 모두 1.8 이상으로 2에 가까운 값임을 확인할 수 있다. 즉, 클린 이미지에 대한 L1 Distance는 0에 가깝고 적대적 예제에 대한 L1 Distance는 2에 가까운 값으로 산출되었다.

Table 3. L1 Distances between Basic Model & Edge Models

eps	Label	Prediction			L1 Distance
		Basic Model	Edge Model		
0.00 (clean)	2	2	Canny	2	0.003679
			Sobel	2	0.003703
			Scharr	2	0.003657
0.02	2	11	Canny	2	1.853495
			Sobel	2	1.853505
			Scharr	2	1.853426
0.05	2	20	Canny	2	1.999081
			Sobel	2	1.998858
			Scharr	2	1.998899
0.1	2	4	Canny	2	1.977092
			Sobel	2	1.808948
			Scharr	2	1.991751
0.2	2	20	Canny	0	1.508391
			Sobel	4	1.399009
			Scharr	4	1.718222
0.3	2	20	Canny	20	1.265619
			Sobel	20	0.457012
			Scharr	4	1.693803

epsilon을 0.2, 0.3으로 설정한 적대적 예제에 대하여, 결국 엣지 모델들도 오답 클래스를 예측하였는데, 이때의 L1 Distance들을 관찰해 보면 대부분 매우 큰 값이 산출되었다. epsilon을 0.3으로 설정한 적대적 예제를 보면, 엣지 모델(Canny, Sobel)이 기본 모델과 동일한 오답 클래스(20)를 예측하였는데 이 경우에도 L1 Distance가 상당히 큰 값임을 확인할 수 있다.

본 논문에서 제시하는 적대적 예제 탐지 방법은 각 모델이 예측한 클래스를 비교하는 것이 아니라 모델 간 L1 Distance를 계산하여 적대적 예제를 탐지하는 것이기 때문에, 비록 엣지 모델이 정답 클래스를 예측하지 못한 경우라도 각 모델 간 L1 Distance의 값이 임계치보다 크다면 적대적 예제를 탐지해 낼 수 있다.

3.2 Deriving threshold

L1 Distance를 판단하기 위한 적정 임계치 도출과 관련하여, 선행 연구[9]에서는 5% 이내로 오탐이 발생할 때의 L1 Distance를 구하여 임계치로 설정했는데, 본 실험을 진행하면서 클린 이미지에 대한 각 모델들의 기본적인 분류 정확도가 고려되지 않는다면 합리적인 임계치 도출이 어려움을 확인하였다. 아래 사례는 기본적인 분류 정확도를 고려하지 않고 오탐율만을 낮추기 위해 임계치를 설정했을 때의 문제점을 보여준다.

Table 4. L1 Distances for a Clean Image

eps	Label	Prediction			L1 Distance
		Basic Model	Edge Model		
0.00 (clean)	11	13	Canny	11	1.697299
			Sobel	11	1.702789
			Scharr	11	1.617076

표 4는 어떤 클린 이미지에 대한 L1 Distance이다. 클린 이미지임에도 기본 모델의 분류 정확도에서 기인된 오답 클래스를 예측하였고, 정답 클래스를 예측한 각 엣지 모델과의 L1 Distance들이 매우 큰 값을 가지고 있다. 만약 오탐율을 줄이기 위해서, 즉, 표 4의 클린 이미지를 적대적 예제로 판단하지 않도록 임계치를 설정해야 한다면 1.702789 이상의 값으로 설정해야 한다.

Table 5. L1 Distances for an Adversarial Example

eps	Label	Prediction			L1 Distance
		Basic Model	Edge Model		
0.02	27	23	Canny	27	1.490360
			Sobel	27	1.506406
			Scharr	27	1.609390

반면, 표 5는 어떤 적대적 예제에 대한 L1 Distance인데, 기본 모델은 오답 클래스를, 엣지 모델들은 정답 클래스를 예측한 전형적인 적대적 예제 탐지 케이스이다. 그런데 가장 큰 L1 Distance가 1.609390이기 때문에, 만약 위에서 도출된 1.702789 이상의 값으로 임계치를 설정하였다면, 결과적으로 이 적대적 예제는 탐지되지 않을 것이다.

이처럼 클린 이미지에 대한 기본적인 분류 정확도에서 기인된 오탐이 있기 때문에, 이를 고려하지 않고 단순히 낮은 오탐율을 기준으로 임계치를 결정할 수는 없다. 본 실험에서는 클린 이미지에 대한 모델들의 분류 정확도(약 92%)를 고려하여 오탐율이 약 8.35%일 때를 기준으로 도출된 L1 Distance인 1.5를 기본 임계치로 설정하였다.

표 6은 클린 이미지 2021개와 적대적 예제(epsilon=0.02) 2021개가 랜덤하게 섞여있는 테스트 데이터셋에 대하여 L1 Distance 임계치를 1.5로 설정했을

때의 적대적 예제에 대한 정답율(탐지율) 및 오답율, 클린 이미지에 대한 정답율 및 오답율(오탐율)이다.

Table 6. Detection Results at Threshold 1.5

eps	Adv. Image		Cln. Image	
	Correct	Wrong	Correct	Wrong
0.02	85.47%	14.53%	91.65%	8.35%

임계치를 1.5 다 더 높게 설정한다면 오탐율은 줄어들지만 그 대신 현재 85.47%인 적대적 예제 탐지율이 낮아지게 될 것이다(테스트 결과, 임계치를 1.6으로 설정했을 때의 탐지율은 81.72%임). 반대로 임계치를 1.5보다 더 낮게 설정한다면 적대적 예제 탐지율은 높아지지만 오탐율이 증가할 것이다(테스트 결과, 임계치를 1.4로 설정했을 때의 오탐율은 9.57%임). 즉, 제시된 임계치는 절대적인 기준이 아니며, 본 탐지 모델을 실제 환경에 적용할 때 필요에 따라 적절히 높이거나 낮춰서 설정할 필요가 있다.

4. Detecting Adversarial Examples

적대적 예제 탐지 테스트는 클린 이미지(2,2021)와 각 epsilon 별 적대적 예제(각 2,021개)들을 섞은 5개의 데이터셋에 대하여 임계치 1.5를 적용하여 수행하였다.

적대적 예제 탐지 시, 입력 이미지가 적대적 예제임에도 일반 모델이 정답 클래스를 맞춘다면 해당 이미지를 적대적 예제가 아닌 클린 이미지로 레이블을 변경하여 결과에 반영하였다. 이것은 적대적 예제의 정의 및 기능적인 측면을 고려하였을 때, 기본 모델로 하여금 오분류를 일으키지 않는 적대적 예제는 해당 모델에 위협이 되지 못하므로 더 이상 적대적 예제가 아니라고 가정을 한 것이다.

Table 7. Detection Results for Adversarial Examples

eps	Adv. Image		Cln. Image	
	Correct	Wrong	Correct	Wrong
0.02	85.47%	14.53%	91.65%	8.35%
	865 Imgs	147 Imgs	2,777 Imgs	253 Imgs
0.05	84.64%	15.36%	90.74%	9.26%
	1,201 Imgs	218 Imgs	2,380 Imgs	243 Imgs
0.1	91.44%	8.56%	88.10%	11.90%
	1,507 Imgs	141 Imgs	2,109 Imgs	285 Imgs
0.2	95.47%	4.53%	88.02%	11.98%
	1,834 Imgs	87 Imgs	1,867 Imgs	254 Imgs
0.3	87.61%	12.39%	88.27%	11.73%
	1,711 Imgs	242 Imgs	1,844 Imgs	245 Imgs

표 7은 임계치 1.5를 기준으로 각 epsilon 별 적대적 예제에 대한 탐지율 및 클린 이미지에 대한 오탐율을 보여주고 있다. 앞서 기술한 것처럼 적대적 예제로 생성되었으나 기본 모델이 정답 클래스를 예측한 이미지는 클린 이미지로 처리되었다(클린 이미지의 개수가 더 많은 이유).

표 7에서 볼 수 있듯이, 전체 적대적 예제들에 대하여 84%~95%의 높은 탐지율을 보이고 있다. 특이사항으로, 노이즈가 작은(epsilon=0.02/0.05) 적대적 예제보다 노이즈가 큰(epsilon=0.1/02) 적대적 예제에 대하여 오히려 탐지율이 더 높았다. 이것은, 약한 노이즈가 적용된 적대적 예제들에 대하여 일반 모델도 약간의 강건성을 유지하여 결과적으로 임계치보다 작은 L1 Distance들이 산출되었고, 강한 노이즈가 적용된 적대적 예제들의 경우 일반 모델은 강건성을 잃지만 엣지 모델들은 여전히 강건성을 유지하여 임계치보다 큰 L1 Distance를 산출했기 때문으로 추측된다.

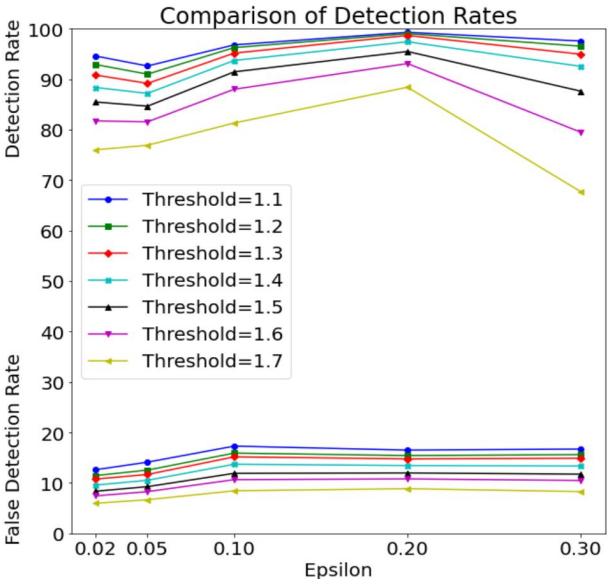

Fig. 8. Comparison of Detection Rates

약한 노이즈의 적대적 예제에 대한 탐지율을 높이기 위해서는 임계치를 낮게 설정해야 하는데, 결과적으로 오탐율이 증가하게 된다. 그림 8은 각 임계치 별 적대적 예제 탐지율과 오탐율을 그래프로 나타낸 것으로서 임계치가 작을수록 탐지율과 오탐율이 높으며, 임계치가 클수록 탐지율과 오탐율이 낮음을 알 수 있다. 탐지율이 높으면서도 오탐율을 낮추는 방안으로, 기본적인 분류 예측 정확도를 향상시키는 방법이 있다. 만약 각 모델들의 분류 예측 정확도를 100%에 가깝게 높인다면, 낮은 오탐율, 그리고 임계치 조정(낮춤)을 통한 높은 탐지율을 기대할 수 있을 것이다.

V. Conclusions

본 연구에서는 객체의 엣지와 분류 모델 간 L1 Distance를 이용하여 적대적 예제를 효과적으로 탐지할 수 있음을 증명하고자 하였다. 먼저 기본 분류 모델과 3종의 엣지 탐지 기법이 적용된 엣지 모델들을 준비하여 30종의 동물 이미지를 대상으로 학습시킨 결과, 엣지 모델의 분류 정확도가 기본 모델과 거의 차이가 없음을 확인하였다. 이어서, 적대적 예제를 생성하여 기본 모델과 엣지 모델의 분류 정확도를 측정한 결과, 기본 모델의 정확도는 매우 낮아졌지만, 엣지 모델은 기본 모델 대비 높은 정확도를 유지하였다. 기본 모델과 엣지 모델 간의 L1 Distance를 측정한 결과, 클린 이미지에 대한 L1 Distance는 작은 값이, 적대적 예제에 대한 L1 Distance는 큰 값이 산출되어, 이를 통한 적대적 예제 탐지 가능성을 확인할 수 있었다. 클린 이미지와 적대적 예제로 구성된 데이터 셋을 대상으로 엣지와 L1 Distance를 이용한 적대적 예제 탐지 모델을 테스트한 결과, 84%~95%의 높은 탐지율을 보여주었다.

결과적으로, 이미지 내 객체의 엣지를 강조하여 학습시킨 엣지 모델이 적대적 예제에 더 강건함과, 그 강건성을 기반으로 기본 모델과의 L1 Distance를 이용하여 적대적 예제를 효과적으로 탐지할 수 있음을 확인할 수 있었다. 엣지 학습을 통한 강건성 향상의 원리는 객체를 더 정확하게 인식하기 위한 정보를 간단히 원본 이미지에 추가하여 학습하는 것으로서, 고비용, 이미지 인식률 저하, 정보 손실 등의 단점이 없어 다양한 분야에서 제약없는 활용이 가능할 것이다. 후속 연구에서는 분류 모델들의 기본적인 분류 정확도를 향상시켰을 때 오탐율의 감소 및 탐지율의 향상 여부를 확인하고자 하며, 이를 위하여 본 연구에서 사용한 분류 모델의 성능을 개선하거나 더 높은 성능의 분류 모델을 선정하여 실험을 진행하고자 한다. 또한 FGSM 이외의 다양한 적대적 예제 생성 기법들에 대해서도 본 탐지 모델의 성능(탐지율)을 확인하고자 한다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00217022).

REFERENCES

[1] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv Preprint arXiv:1412.6572, Dec. 2014, DOI: 10.48550/arXiv.1412.6572

[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust physical-world attacks on deep learning visual classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625-1634, Jun. 2018, DOI: 10.1109/cvpr.2018.00175

[3] A. Liu, J. Guo, J. Wang, S. Liang, R. Tao, W. Zhou, C. Liu, X. Liu and D. Tao, "X-adv: Physical adversarial object attacks against x-ray prohibited item detection," 32nd USENIX Security Symposium (USENIX Security 23), pp. 3781-3798, Aug. 2023.

[4] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," Pattern Recognition, vol. 110, pp. 107332, Feb. 2021, DOI: 10.1016/j.patcog.2020.107332

[5] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," IEEE Access, vol. 6, pp. 14410-14430, 2018, DOI: 10.1109/access.2018.2807385

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards deep learning models resistant to adversarial attacks," International Conference on Learning Representations, Feb. 2018.

[7] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry, "Robustness may be at odds with accuracy," International Conference on Learning Representations, May 2019.

[8] A. Borji, "Shape defense," in I (Still) can'T Believe it's Not Better! Workshop at NeurIPS 2021, pp. 15-20, Feb. 2022.

[9] W. Xu, D. Evans and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," NDSS Symposium, 2018, DOI: 10.14722/ndss.2018.23198

[10] S. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582, Jun. 2016, DOI: 10.1109/cvpr.2016.282

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 IEEE Symposium on Security and Privacy (Sp), pp. 39-57, May 2017, DOI: 10.1109/sp.2017.49

[12] D. Meng and H. Chen, "Magnet: A two-pronged defense against adversarial examples," in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135-147, Oct. 2017, DOI: 10.1145/3133956.3134057

[13] J. H. Metzen, T. Genewein, V. Fischer and B. Bischoff, "On detecting adversarial perturbations," International Conference on Learning Representations, 2017.

[14] L. G. Roberts, "Machine Perception of Three-Dimensional Solids," PhD Thesis. Massachusetts Institute of Technology, 1963.
[15] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," a talk at the Stanford Artificial Project in, pp. 271-272, 1968.
[16] J. Canny, "A computational approach to edge detection," Readings in Computer Vision, pp. 184-203, 1987, DOI: 10.1016/b978-0-08-051581-6.50024-6
[17] H. Scharr, "Optimal operators in digital image processing," http://www.ub.uni-heidelberg.de/archiv/962, 2000.
[18] J. Bright, "ANIMALS(30 Animal Species for Easy train)," https://www.kaggle.com/datasets/jerrinbright/cheetahtigerwolf

Authors

Jaesung Shim received the M.S. degree in Information Security from Soongsil University, in 2008. He is currently a Ph.D. candidate in Computer Engineering at Chungbuk National University.

He is interested in computer vision and information security.

Kyuri Jo received her B.S. and Ph.D. degree in Computer Science and Engineering from Seoul National University in 2013 and 2018, respectively and worked as a postdoctoral researcher at Bio and Helath Informatics Lab., Seoul National University. She is currently an associate professor in the Department of Computer Engineering at Chungbuk National University since September 2019. Her current research interests include artificial intelligence, machine learning and bioinformatics.