Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.29, NO.6, Dec. 2019 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2019.29.6.1383

감쇠 요소가 적용된 데이터 어그멘테이션을 이용한 대체 모델 학습과 적대적 데이터 생성 방법

민 정 기,† 문 종 섭‡
고려대학교 정보보호대학원

A Substitute Model Learning Method Using Data Augmentation with a Decay Factor and Adversarial Data Generation Using Substitute Model

Jungki Min,† Jong-sub Moon‡
Graduate School of Information Security, Korea University

요 약

적대적 공격은 기계학습 분류 모델의 오분류를 유도하는 적대적 데이터를 생성하는 공격으로, 실생활에 적용된 분류 모델에 혼란을 야기하여 심각한 피해를 발생시킬 수 있다. 이러한 적대적 공격 중 블랙박스 방식의 공격은, 대상 모델과 유사한 대체 모델을 학습시켜 대체 모델을 이용해 적대적 데이터를 생성하는 공격 방식이다. 이 때 사용되는 야코비 행렬 기반의 데이터 어그멘테이션 기법은 합성되는 데이터의 왜곡이 심해진다는 단점이 있다. 본 논문은 기존의 데이터 어그멘테이션 방식에 존재하는 단점을 보완하기 위해 감쇠 요소를 추가한 데이터 어그멘테이션을 사용하여 대체 모델을 학습시키고, 이를 이용해 적대적 데이터를 생성하는 방안을 제안한다. 실험을 통해, 기존의 연구 결과보다 공격 성공률이 최대 8.5% 가량 높음을 입증하였다.

ABSTRACT

Adversarial attack, which geneartes adversarial data to make target model misclassify the input data, is able to confuse real life applications of classification models and cause severe damage to the classification system. An Black-box adversarial attack learns a substitute model, which have similar decision boundary to the target model, and then generates adversarial data with the substitute model. Jacobian-based data augmentation is used to synthesize the training data to learn substitutes, but has a drawback that the data synthesized by the augmentation get distorted more and more as the training loop proceeds. We suggest data augmentation with 'decay factor' to alleviate this problem. The result shows that attack success rate of our method is higher(around 8.5%) than the existing method.

Keywords: Deep Learning, Adversarial Data Generation, Data Augmentation

I. 서 론

기계학습 모델 중에서 분류 모델은 이미지 인식, 영상 인식, 음성 인식과 같은 다양한 분야에서 활용되고 있는 모델이다. 이러한 분류 모델의 입력 데이터에 아주 미세한, 인간은 구별이 불가능한 오차를 주입하여 분류 모델의 오분류(misclassification)를 유도하는 적대적 공격(Adversarial Attack)에 대한 연구가 활발히 진행되고 있다[1, 2, 4]. 아래 Fig. 1.은 이러한 적대적 공격의 예시이다.

인간의 눈으로 봤을 때 Fig. 1.의 두 이미지는 동일한 좌회전 표지판으로 보인다. 그러나 인공 신경망(neural network)을 활용해 학습된 교통 표지판

Received(10. 11. 2019), Accepted(11. 5. 2019)
† 주저자, eternalray@korea.ac.kr
‡ 교신저자, jsmoon@korea.ac.kr(Corresponding author)

Fig. 1. Traffic sign images of left turn. Left image is an original image. Right image is an image generated using adversarial attack

분류 모델은 좌측의 원본 이미지를 좌회전 표지판으로 제대로 분류하는 반면, 우측의 적대적 공격에 의해 생성된 이미지를 속도 제한 표지판으로 오분류 하게 된다. 악의적인 사용자는 적대적 공격을 통해 생성된 이미지를 이용해 자율 주행 자동차에 탑재된 교통 표지판 분류 모델에 혼란을 야기할 수 있고, 이는 심각한 인명 피해로 연결 될 수 있다. 이와 같은 예시뿐만 아니라 적대적 공격은 분류 모델을 활용하는 분야 전반에 영향을 줄 수 있기 때문에, 적대적 공격 방법과 이에 대한 해결 방안 연구의 중요성이 높아지고 있다.

적대적 공격은 본래 공격 대상 모델의 내부 파라미터를 알고 있는 상태에서 수행하는 공격이다. 이러한 이유로, 공격의 실효성에 대한 의문이 제기 되었으나, 공격 대상 모델의 내부 파라미터를 모르는 상태에서 수행할 수 있는 블랙박스 공격 방안이 제안되었다[3]. 이는 대상 모델과 비슷한 결정 경계를 갖는 대체 모델(substitute model)을 학습시켜, 이 대체 모델의 파라미터를 활용하여 공격을 수행하는 방법이다.

기존의 [3]에서 사용된 야코비 행렬 기반 데이터 어그멘테이션 (Jacobian-based Data Augmentation)은 데이터 어그멘테이션이 진행 될수록 데이터의 왜곡이 점점 심해진다는 단점이 있다. 때문에 심하게 왜곡된 데이터들을 사용하여 대체 모델을 학습하는 경우 오버 피팅이 발생하게 되어 대체 모델과 대상 모델의 결정 경계(decision boundary)가 상이해 질 수 있다.

본 논문은 대체 모델 학습 시에 사용되는 야코비 행렬 기반 데이터 어그멘테이션의 문제점을 보완하여 공격 성공률을 높일 수 있는 새로운 방안을 제안한다. 본 논문에서는 데이터 어그멘테이션 과정에서 생성되는 데이터의 왜곡되는 정도를 조절하기 위한 감쇠 요소(decay factor)를 추가하여 데이터의 왜곡되는 정도를 감소시켰고, 이를 통해 공격 성공률이 더 높아짐을 Modified National Institute of Standards and Technology(MNIST) 데이터셋과 German Traffic Sign Recognition Benchmark(GTSRB) 데이터셋에 대한 실험을 통해 입증보인다.

본 논문의 구성은 다음과 같다. 2장에서는 분류 모델에 대한 대표적인 적대적 공격 방법을 화이트박스 공격과 블랙박스 공격으로 나누어 소개한다. 3장에서는 [3]에서 제시된 블랙박스 공격 기법에 감쇠 요소를 적용함으로써 데이터의 왜곡을 줄일 수 있고, 이에 따라 공격 성공률을 높일 수 있음을 보인다. 4장에서는 3장에서 제안된 방법을 통한 실험 결과를 보이고 마지막 5장에서는 결론과 앞으로의 연구 방향을 제시한다.

II. 관련 연구

2.1 화이트박스 공격

화이트박스 공격[5, 6 & 7]은 대상 모델에 대한 파라미터 정보를 알고 있는 상태에서 대상 모델의 오분류를 유도하는 적대적 데이터를 생성하는 공격이다. 화이트박스 공격 방식은 크게 원본 데이터 x_i의 실제 클래스 라벨에 해당하는 y_i를 제외한 어떠한 클래스로든 오분류가 되도 상관 없는 타겟 미지정 공격(non-targeted attack)[7]과, 특정 클래스 라벨 $\hat{y}_i(\hat{y}_i \neq y_i)$로의 오분류를 유도하는 타겟 지정 공격(targeted attack)[5, 6]으로 나뉜다.

2.1.1 타겟 지정 공격

타겟 지정 공격의 대표적인 예로는 특정 클래스 라벨에 대한 모델의 야코비 행렬을 이용하는 Jacobian-based Saliency Map Attack(JSMA)[5]와 L2 거리를 이용하는 C&W Attack[6]이 있다. 타겟 지정 공격은 특정 클래스로의 유도를 가능하게 한다는 장점이 있지만, 이를 위한 연산이 타겟 미지정 공격에 비해 월등히 많이 필요하기 때문에 실용성이 떨어진다는 단점이 있다[3, 8].

2.1.2 타겟 미지정 공격

타겟 미지정 공격의 대표적인 예로는 Fast Gra

dient Sign Method(FGSM)[7]이 있다. FGSM은 모델 F와 비용 함수 $c(F,\vec{x},y)$가 주어 졌을 때, 적대적 데이터 $\vec{x^*}=\vec{x}+\vec{\delta_x}$를 생성하기 위한 오차값 $\vec{\delta_x}$를 다음과 같이 계산한다.

$$\vec{\delta_x}=\epsilon sgn(\nabla_{\vec{x}}c(F,\vec{x},y)) \quad (1)$$

이 때 $sgn(\nabla_{\vec{x}}c(F,\vec{x},y))$은 모델의 비용 함수의 경사값(gradient)의 부호(+1 또는 -1)이다. 원래 모델 학습 시에는 비용 함수의 값을 최소화하기 위해 경사값의 반대 부호를 파라미터에 더하는 경사 하강법(gradient descent)[9]을 사용하지만, FGSM은 모델의 오분류를 유도하려는 목적으로 비용 함수의 값을 최대화하기 위해 경사값의 부호를 그대로 입력값에 더한다. 이 때 ϵ은 더해지는 오차의 크기를 조절하기 위한 값으로 ϵ이 커질수록 모델 F가 $\vec{x^*}$을 오분류할 확률이 높아지지만, 그만큼 데이터의 왜곡되는 정도가 심해지기 때문에 인간이 탐지하기 쉬워진다는 단점이 있다.

2.2 블랙박스 공격

화이트박스 공격의 경우, 적대적 데이터를 생성하기 위한 오차값 계산 시에 대상 모델의 내부 정보를 알 필요가 있다. 따라서 실제 상황에서는 악의적인 사용자가 대상 모델의 구조를 상세히 알고, 대상 모델의 파라미터에 접근할 수 있어야 한다는 강력한 가정이 필요하므로 화이트박스 공격은 실용성이 떨어진다는 한계점이 존재한다[3]. 따라서 이러한 한계점을 극복하기 위해 블랙박스 공격 방안이 제안되었다[3]. 블랙박스 공격은 대상 모델과 유사한 결정 경계를 갖는 대체 모델을 학습시켜, 대체 모델의 내부 정보를 이용해 적대적 데이터를 생성하고, 이를 대상 모델을 공격하는데 사용하는 공격 방법이다. 기존 연구 결과[7, 10 & 15]를 통해 대상 모델과 동일한 분포를 갖는 데이터셋을 이용해 학습시킨 대체 모델은 대상 모델과 유사한 내부 정보를 갖는 것이 확인되었다. 이러한 성질을 전이성(transferability)이라고 하며, 이는 대체 모델의 파라미터 정보를 이용해 생성된 적대적 데이터는 대체 모델에서 오분류 될 뿐만 아니라 매우 높은 확률로 대상 모델에서도 오분류 됨을 의미한다.

대체 모델을 학습하기 위해서는 학습 데이터가 필요한데, 악의적인 사용자가 대상 모델을 학습시키는데 사용되는 실제 데이터에 접근할 수 있다는 가정은 현실적으로 무리가 있다. 따라서 대체 모델 학습 데이터를 확보하기 위해, 적은 양의 초기 데이터셋에 대해 데이터 어그멘테이션을 수행한다. [3]에서는 야코비 행렬 기반의 데이터 어그멘테이션을 수행하여 트레이닝 데이터를 확보하였다. 야코비 행렬 기반의 데이터 어그멘테이션은 다음과 같이 수행 된다.

$$S_{\rho+1}=\{\vec{x}+\lambda sgn(J_F[\tilde{O}(\vec{x})]):\vec{x}\in S_\rho\}\cup S_\rho \quad (2)$$

이 때 ρ는 어그멘테이션이 진행 되는 반복 횟수,

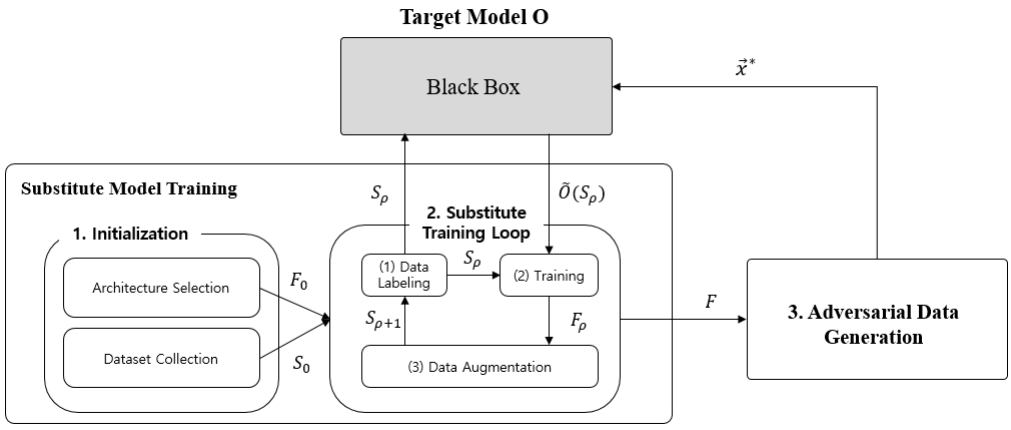

Fig. 2. Process of Black-box Adversarial Attack

S는 데이터셋, J_F는 모델 F에 대한 야코비 행렬이고, $\tilde{O}(\vec{x})$는 입력 데이터 $\vec{x}$에 대한 대상 모델의 분류 결과이다. 따라서 $sgn(J_F[\tilde{O}(\vec{x})])$는 야코비 행렬 중 대상 모델의 분류 결과에 해당하는 클래스 라벨에 대한 경사값의 부호라고 할 수 있다. 이를 크기를 조절하는 λ 상수를 곱해 입력 데이터 $\vec{x}$에 더하여 새로운 데이터를 생성한다.

III. 제안 방법

3.1 공격 개요

본 논문에서는 감쇠 요소를 적용한 야코비 행렬 기반의 데이터 어그멘테이션을 통해 생성된 데이터셋을 대체 모델을 학습시키는데 사용하는 새로운 블랙박스 공격 방안을 제안한다.

대상 모델 O는 DNN을 이용한 분류 모델로, 입력 데이터가 $\vec{x}$가 주어질 시에 해당 입력 데이터가 속할 확률이 가장 높은 클래스에 대한 라벨을 출력한다. 이는 다음과 같은 식으로 표현할 수 있다. 이때 $O_j(\vec{x})$는 클래스 j에 대한 확률이다.

$$\tilde{O}(\vec{x}) = \arg\max_{j \in 0..N-1} O_j(\vec{x}) \quad (3)$$

악의적인 사용자는 대상 모델에 데이터를 입력했을 때 반환 되는 $\tilde{O}(\vec{x})$ 이외에, 대상 모델의 내부 구조, 파라미터 정보, 트레이닝 데이터와 같은 어떠한 정보에도 접근할 수 없다.

악의적인 사용자의 목표는 대상 모델 O의 오분류를 유도하는 적대적 데이터 $\vec{x^*}$을 생성하는 것이다. $\vec{x^*}$은 다음과 같은 식을 만족한다.

$$\vec{x^*} = \vec{x} + \arg\min\{\vec{z}: \tilde{O}(\vec{x}+\vec{z}) \neq \tilde{O}(\vec{x})\} \quad (4)$$

식(4)의 의미는 입력 데이터에 더해질 오차 $\vec{z}$는 최소한이어야 한다는 것이다. 오차가 크면 클수록 대상 모델의 오분류 확률은 증가하지만, 인간이 데이터의 왜곡을 감지할 수 있는 확률도 같이 증가하게 된다. 따라서 오차는 인간이 감지할 수 없을 정도의 미세한 작은 크기를 유지하면서, 대상 모델의 오분류를 유도할 정도로는 커야 한다.

제안하는 공격 방안의 전체적인 흐름은 Fig. 2.와 같다. 공격 과정은 크게 대체 모델 학습(Substitute Model Training)과 적대적 데이터 생성(Adversarial Data Generation)으로 나뉜다. 대체 모델 학습은 대체 모델의 구조를 정하고 초기 데이터셋 S_0를 확보하는 초기화 단계(Fig. 2.의 1.)와 데이터 어그멘테이션을 통해 트레이닝 데이터 S_ρ를 생성하고 대체 모델을 학습하는 대체 학습 반복(Fig. 2.의 2.)으로 진행된다. 대체 모델 학습이 완료되면, 학습된 대체 모델 F를 이용하여 적대적 데이터 $\vec{x^*}$를 생성한다.(Fig. 2.의 3.)

3.2 대체 모델 학습

Fig. 2.의 대체 모델 학습에 대한 알고리즘은 Table 1.과 같다. 먼저 악의적인 사용자는 학습에 사용 될 초기 데이터셋 S_0를 확보해야 한다. 악의적인

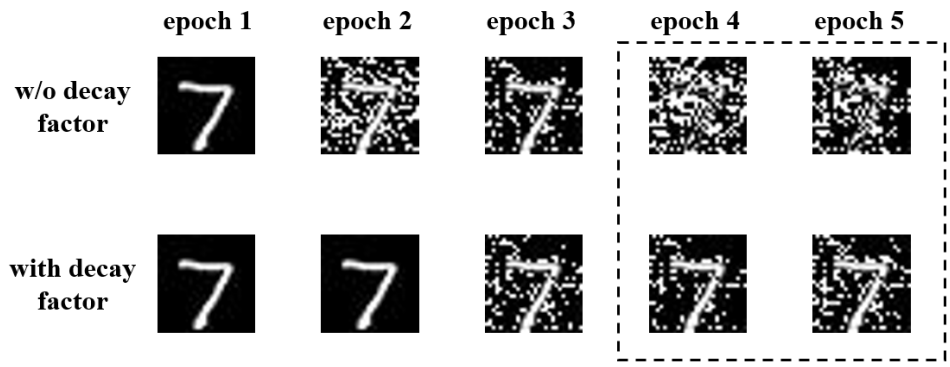

Fig. 3. The result of Jacobian-based data augmentation per epoch

사용자는 대상 모델의 학습 데이터에 접근할 수 없기 때문에, 미리 알고 있는 정보를 바탕으로 데이터를 수집할 수밖에 없다. 예를 들어 대상 모델이 동물 이미지 분류 모델일 경우, 악의적인 사용자는 해당하는 동물 이미지를 모아서 대체 모델을 학습시킬 초기 데이터 셋으로 사용해야 한다.

다음으로 악의적인 사용자는 대체 모델 F의 구조를 결정한다.(Table 1.의 line 1.) 대상 모델이 특정 데이터(ex. 이미지, 음성, 영상 등)에 대한 분류 모델이라는 점을 바탕으로, 해당 데이터의 특성에 맞는 대체 모델의 구조를 정한다. 예를 들어 이미지 데이터의 경우에는 CNN[11, 12], 음성 데이터의 경우에는 RNN[13, 14]을 사용하는 식으로 대체 모델을 정할 수 있다. 뿐만 아니라, 대체 모델의 레이어나 하이퍼 파라미터들에 대한 구성 역시 자유롭게 정하면 된다. 악의적인 사용자는 추후에 다양한 구조의 대체 모델들을 실험해 보고, 가장 공격 성공률이 높은 모델을 사용하면 된다. 대체 모델은 꼭 딥 러닝 모델일 필요는 없으나[15] 본 논문에서는 딥 러닝 모델로 한정 지었다.

대체 모델을 학습시키기 위해, 악의적인 사용자는 우선 S_ρ에 대한 데이터 라벨링을 수행한다.(line 4.) 이 때 대상 모델 O에게 각 데이터에 대한 분류 결과를 질의하고, 이를 데이터의 실제 라벨로 여기고 대체 모델 학습을 수행한다.(line 6.) 그 후 대체 모델의 트레이닝 데이터의 수를 증가시키기 위해, 야코비 행렬 기반의 데이터 어그멘테이션을 수행한다.(line 9.) 이 때 $\lambda_{\rho+1}$이 λ_ρ보다 작도록 하는 감쇠 요소 d_ρ를 적용한다.(line 10.)

Table 1. Pseudo Code for Substitute Model Training Algorithm

Input: O, $\max_\rho$, S_0, λ_0
1: Select model F
2: for $\rho \in 0..\max_\rho - 1$ do
3: // Data labeling
4: $D = \{(\vec{x}, \widetilde{O}(\vec{x}) : \vec{x} \in S_\rho\}$
5: // Training
6: $\theta_F = train(F, D)$
7: // Jacobian-based data augmentation
8: // with decay factor
9: $S_{\rho+1} = \{\vec{x} + \lambda_\rho sgn(J_F[\widetilde{O}(\vec{x})]) : \vec{x} \in S_\rho\} \cup S_\rho$
10: $\lambda_{\rho+1} = \lambda_\rho \times d_\rho (0 < d_\rho < 1)$
11:end for
12:return θ_F

3.3 감쇠 요소를 적용한 데이터 어그멘테이션

[3]에서 제안한 방안과 본 논문의 다른 점은 바로 감쇠 요소에 있다. 기존의 야코비 행렬 기반의 데이터 어그멘테이션의 경우, 데이터 어그멘테이션이 진행 될수록 데이터의 왜곡이 점점 심해지기 때문에 원본 데이터와 생성된 데이터간의 연관성이 빠르게 감소하게 된다는 단점이 있다. 이렇게 왜곡이 심한 데이터를 사용해 대체 모델을 학습할 경우, 왜곡된 데이터에 대해 오버 피팅이 발생하게 되고, 이는 대상 모델과 유사한 결정 경계를 갖도록 대체 모델을 학습시키는 목적에 방해된다. 따라서 데이터의 다양성은 증가시키면서 왜곡 되는 정도를 줄일 수 있다면, 대체 모델이 대상 모델의 결정 경계를 더 잘 학습할 수 있게 된다. Fig. 3.은 감쇠 요소의 적용 유무에 따라 생성된 데이터의 왜곡 정도가 다름을 보여주는 그림이다. 감쇠 요소를 적용하지 않은 경우, 학습 후반부에 생성된 데이터들은 원본 데이터의 형태를 거의 갖고 있지 않다. 반면 감쇠 요소를 적용한 경우, 학습 후반부에 생성된 데이터들도 원본 데이터와 유사한 형태를 유지하고 있다. 이는 감쇠 요소를 적용할 경우, 데이터 어그멘테이션을 통해 대체 모델을 학습시킬 충분한 데이터를 확보할 수 있으면서, 심하게 왜곡된 데이터에 대한 오버 피팅이 발생하지 않고 대체 모델이 대상 모델의 결정 경계를 잘 학습할 수 있음을 의미한다. 감쇠 요소 d_ρ를 구하는 식은 다음과 같다.

$$d_{\rho+1} = \begin{cases} d_\rho \\ \dfrac{1}{1+e^{-\lambda_\rho \times t}} \end{cases} \quad (5)$$

d_ρ를 구하는 방법은 크게 두 가지가 있다. 첫 번째 방법은 d_ρ를 상수로 사용하는 것이다. 이는 매 반복마다 고정된 비율만큼 λ_ρ를 감쇠 시킴을 의미한다. 두 번째 방법은 시그모이드 함수를 사용하는 것이다. 시그모이드 함수는 입력 값을 0과 1사이의 값으로 변환해 준다. 이 때 입력 값이 0에 가까울수록 λ_ρ가 감쇠되는 비율이 증가하기 때문에, 초기 λ_0값에 따라 이를 보정해주기 위한 상수 t가 곱해진다. 상수 t는 학습 초반부에는 λ_ρ의 크기를 거의 감소시키지

않다가, 학습 후반부에 λ_ρ의 크기를 크게 감소시켜 데이터가 왜곡되는 정도를 크게 줄이는 역할을 한다.

3.4 적대적 데이터 생성

대체 모델 F를 학습 한 뒤, 악의적인 사용자는 F의 내부 파라미터 정보를 이용하여 적대적 데이터 $\overrightarrow{x^*}$를 생성한다. 본 논문에서는 벤치마크로 흔히 사용되는 FGSM을 사용하여 적대적 데이터를 생성했다. 악의적인 사용자는 반복적인 적대적 데이터 생성을 통해, 대상 모델 O가 생성된 데이터를 오분류 할 확률은 높이면서, 생성된 데이터가 인간의 탐지로는 구별할 수 없도록 하는 최적의 ϵ을 찾는다.

IV. 실험 및 평가

4.1 실험 환경 및 실험 데이터

Table 2.는 실험이 이루어진 시스템의 환경과 사용한 딥 러닝 라이브러리 TensorFlow[16]의 버전이다.

실험에 사용한 데이터셋은 [3]에서와 마찬가지로 총 두 종으로, MNIST 데이터셋과 GTSRB 데이터셋을 사용했다. MNIST 데이터셋은 0에서 9까지의 숫자를 표현한 28x28 크기의 흑백 이미지이다. MNIST 데이터셋은 60,000개의 학습 데이터와 10,000개의 테스트 데이터로 구성되어 있으며, 10,000개의 테스트 데이터중 150개를 대체 모델을 학습하는데 사용할 초기 데이터셋으로 설정하고 실험을 진행했다.

GTSRB 데이터셋은 독일 교통 표지판 이미지로, 43종의 교통 표지판으로 구성되어 있다. GTSRB 이미지는 크기가 제각각이고 RGB 인코딩되어 있기 때문에, 학습을 위해 이미지를 32x32 크기로 재조정했고, RGB 인코딩을 흑백으로 변환하여 사용했다. GTSRB 데이터셋은 39,209개의 학습 데이터셋과 12,630개의 테스트 데이터로 구성되어 있으며, 12,630개의 테스트 데이터중 1000개를 대체 모델 학습 시의 초기 데이터셋으로 설정하고 실험을 진행했다. MNIST 데이터셋과 초기 학습 데이터셋 수의 차이가 있는 이유는 GTSRB 이미지가 더 복잡하기 때문에 원활한 학습을 위해 초기 데이터셋의 수를 크게 설정했다.

Table 2. Experiment Environment

OS	Ubuntu 18.04
CPU	Intel Core i7-4790
RAM	16GB
TensorFlow	1.13.1

4.2 실험 구성

실험은 MNIST 데이터셋과 GTSRB 데이터셋에 대한 분류 모델을 대상으로 하는 적대적 공격을 수행하고, 이에 대한 공격 성공률을 측정하는 방식으로 진행 했다. 대상 모델은 공격을 위해 미리 학습시켜 놓은 DNN 모델로, 모델의 테스트 정확도는 MNIST 분류 모델의 경우 약 99.2%, GTSRB 분류 모델의 경우 약 97.3% 이다. 실험을 통해 본 논문에서 감쇠 요소를 적용한 블랙 박스 공격과, [3]에서 제안한 기존의 공격 기법간의 공격 성공률을 비교하였다. 공격 성공률이란, 적대적 데이터를 대상 모델이 오분류 했을 시 공격이 성공한 것으로 간주하며, 이때 공격 성공률은 다음과 같이 계산한다.

$$\text{sucess} \quad \text{rate} = (1 - \text{accuracy}) \times 100 \tag{6}$$

적대적 데이터 생성에 사용된 원본 데이터들은 각 데이터셋의 테스트 데이터를 사용하였다. 실제로는 악의적인 사용자가 모든 테스트 데이터에 접근하지 못할 수도 있지만, 본 논문에서는 많은 데이터에 대한 공격 성공률을 측정하기 위해 테스트 데이터 전체를 사용하였다. 각 데이터셋 당 대체 모델 두 종류를 선정하여 실험을 진행 하였으며, 대체 모델의 자세한 구조는 Table. 3.에 명시되어 있다.

대체 모델 학습은 총 6번의 어그멘테이션 반복으로 진행되며, 어그멘테이션 반복 당 데이터 어그멘테이션이 수행되어 데이터가 두 배씩 늘어난다. 이 때 $\lambda_0 = 0.1$이며, 상수 감쇠 요소 인 경우에 $d_\rho = 0.95$이고, 시그모이드 감쇠 요소 인 경우에 λ_ρ에 곱해지는 상수 $t = 30$이다. 또한 각 어그멘테이션 반복 당 10번의 트레이닝 반복이 수행된다. 학습은 SGD (Stochastic Gradient Descent) 알고리즘[17]을 사용하였다. 이 때 배치 크기는 MNIST의 경우 128, GTSRB의 경우 32이고, 러닝 레이트는 10^{-2}이며, Adam 최적화 알고리즘[18]을 사용하였다. Gao 등[19]은 대체 모델 학습 시 앙상블 학습[20]을 사용하였고, 실험을 통해 타겟 지정 공격을 사용

Table 3. Architectures of Substitute Models(Conv: Convolution layer, Relu: Relu activation, FC:Fully-connected layer)

MNIST		GTSRB	
A	B	C	D
Conv(32,2)+Relu MaxPooling(2) Conv(64,2)+Relu MaxPooling(2) FC(200)+Relu FC(200)+Relu FC(10)+Softmax	Conv(32,2)+Relu MaxPooling(2) Conv(64,2)+Relu MaxPooing(2) FC(10)+Softmax	Conv(32,3)+Relu Conv(32,3)+Relu MaxPooling(2) Conv(64,3)+Relu Conv(64,3)+Relu MaxPooling(2) Conv(128,3)+Relu Conv(128,3)+Relu MaxPooling(2) FC(512)+Relu FC(43)+Softmax	Conv(32,3)+Relu Conv(32,3)+Relu MaxPooling(2) Conv(64,3)+Relu Conv(64,3)+Relu MaxPooling(2) Conv(128,3)+Relu Conv(128,3)+Relu MaxPooling(2) FC(43)+Softmax

하는 경우 공격 성공률이 증가함을 보였다. 그러나 타겟 미지정 공격에 대해서는 앙상블 학습이 효과가 없음이 Liu 등[21]에 의해 알려져 있다. 따라서 본 논문에서는 앙상블 학습에 대한 추가적인 실험을 진행하지 않았다.

적대적 데이터 생성에는 FGSM을 사용하였으며, ϵ의 값을 다르게 하며 생성된 적대적 데이터에 대한 공격 성공률을 측정하였다.

4.3 실험 결과

Table 4.는 MNIST 데이터셋에 대하여 대체 모델 A와 B를 학습시켜 적대적 공격을 수행하였을 때, 감쇠 요소의 유무에 따른 공격 성공률을 비교한 표이다. ϵ의 크기가 커짐에 따라 감쇠 요소의 적용 유무가 공격 성공률에 큰 영향을 끼치는 것으로 보인다. 특히 $\epsilon = 0.3$인 경우에 이 차이가 극명하게 드러나는데, 대체 모델 A의 경우, 감쇠 요소를 적용하지 않았을 때의 공격 성공률은 약 40.46%인 반면, 상수 감쇠 요소를 적용했을 때의 공격 성공률은 약 49.14%로, 약 8.5% 가량 차이가 난다. 대체 모델 B의 경우, 감쇠 요소를 적용하지 않았을 때의 공격 성공률은 약 40.58%인 반면, 시그모이드 감쇠 요소를 적용했을 때의 공격 성공률은 약 47.35%로, 약 7%가량 차이가 난다. 이외에도 전반적으로 감쇠 요소를 적용하였을 때가 더 공격 성공률이 높음을 알 수 있다.

Table 5.는 GTSRB 데이터셋에 대하여 대체 모델 C와 D를 학습 시켜 적대적 공격을 수행하였을 때, 감쇠 요소의 유무에 따른 공격 성공률을 비교한 표이다. MNIST 데이터셋에 대한 실험과 마찬가지로 전반적으로 감쇠 요소를 적용하였을 때의 공격 성공률이 더 높음을 알 수 있다. 감쇠 요소를 적용한 경우, 약 1~2% 가량 공격 성공률이 높음을 확인할 수 있다. 이를 통해 본 논문에서 제안한 공격 방법이 [3]보다 우수함을 알 수 있다.

Table 4. Comparison of success rates when using MNIST dataset(w/o d: without decay factor, constant d: constant decay factor, sigmoid d: sigmoid decay factor)

ϵ	success rate(%)					
	substitute model A			substitute model B		
	w/o d	constant d	sigmoid d	w/o d	constant d	sigmoid d
0.1	1.31	**1.28**	**1.36**	1.36	**1.39**	**1.33**
0.2	6.85	**8.15**	**7.02**	6.42	**7.06**	**7.59**
0.3	40.46	**49.14**	**45.83**	40.58	**43.27**	**47.35**
0.4	71.52	**75.31**	**76.45**	68.75	**72.64**	**73.49**
0.5	83.06	**85.18**	**85.83**	81.22	**83.75**	**83.95**

Table 5. Comparison of success rates when using GTSRB dataset(w/o d: without decay factor, constant d: constant decay factor, sigmoid d: sigmoid decay factor)

ϵ	success rate(%)					
	substitute model C			substitute model D		
	w/o d	constant d	sigmoid d	w/o d	constant d	sigmoid d
0.05	33.78	**35.33**	**34.06**	31.54	**33.45**	**33.34**
0.1	56.81	**58.28**	**56.96**	54.39	**55.83**	**56.53**
0.15	71.07	**71.98**	**71.45**	70.43	**69.88**	**69.75**

V. 결 론

본 논문에서는 야코비 행렬 기반의 데이터 어그멘테이션 기법에 감쇠 요소를 적용하여 대체 모델을 학습시키고, 대체 모델을 이용해 대상 분류 모델에 대한 적대적 공격을 수행하는 블랙박스 공격 방안을 제시하였다. 감쇠 요소를 적용하여 데이터 어그멘테이션 과정에서 생성되는 데이터의 왜곡 정도를 줄여, 대체 모델을 더 잘 학습할 수 있도록 하였고, 실험을 통해 기존의 공격 방안보다 공격 성공률이 더 높음을 입증하였다.

향후 연구를 통해 본문에서 사용된 FGSM과 같은 타겟 미지정 공격이 아닌 타겟 지정 공격을 사용하는 블랙박스 공격에 대해서도 감쇠 요소를 적용하는 연구가 가능하고, 감쇠 요소의 수학적인 의미를 더 분석하고 이를 일반화하여 더욱 다양한 모델에 적용하는 연구를 통해 제시한 공격 방안을 발전시킬 수 있을 것이다.

References

[1] L. Huang, A. D. Joseph, and B. Nelson, "Adversarial machine learning." In Proceedings of the 4th ACM workshop on Security and artificial intelligence, pp. 43-58, October 2011.

[2] B. Biggio, I. Corona, and D. Maiorca, "Evasion attacks against machine learning at test time." Joint European conference on machine learning and knowledge discovery in databases, pp. 387-402, September 2013.

[3] N. Papernot, P. McDaniel, and I. Goodfellow, "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506-519, April 2017.

[4] S. Qiu, Q. Liu, and S. Zhou, "Review of artificial intelligence adversarial attack and defense technologies." Applied Sciences, vol. 9, no. 5, pp. 909-938, March 2019.

[5] N. Papernot, P. McDaniel, and S. Jha, "The limitations of deep learning in adversarial settings." 2016 IEEE European Symposium on Security and Privacy, pp. 372-387, March 2016.

[6] N. Carlini, and D. Wagner, "Towards evaluating the robustness of neural networks." 2017 IEEE Symposium on Security and Privacy, pp. 39-57, May 2017.

[7] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples." arXiv preprint arXiv: 1412.6572, 2014.

[8] X. Yuan, P. He, and Q. Zhu, "Adversarial examples: attacks and defenses for deep learning." IEEE transactions on neural networks and learning systems, vol.9, no.5, pp. 2805-2824, January 2019.

[9] D. Bertsekas, "Nonlinear programming." Journal of the Operational Research Society, vol.48, no.3, pp. 334, 1997.

[10] C. Szegedy, W. Zaremba, and I. Sutskever, "Intriguing properites of neural networks." arXiv preprint arXiv:1312.6199, 2013.

[11] Y. Lecun, P. Haffner, and L. Bottou, "Object recognition with gradient-based learning." Shape, contour and grouping in computer vision, Springer, Berlin, Heidelberg, pp. 319-345, 1999.

[12] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems, vol.1, pp. 1097-1105, December, 2012.

[13] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition." arXiv preprint arXiv:1402.1128, 2014.

[14] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." Fifteenth annual conference of the international speech communication association, pp. 338-342, January 2014.

[15] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning from phenomena to black-box attacks using adversarial samples." arXiv preprint arXiv:1605.07277, 2016.

[16] M. Abadi, P. Barham, and J. Chen, "Tensorlfow: A system for large-scale machine learning." 12th USENIX Symposium on Operating Systems Design and Implementation, pp. 265-283, November 2016.

[17] L. Bottou, "Large-scale machine learning with stochastic gradient descent." Proceedings of COMPSTAT'2010. Physica-Verlag HD, pp. 177-186, August 2010.

[18] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

[19] X. Gao, Y. Tan, and H. Jiang, "Boosting targeted black-box attacks via ensemble substitute training and linear augmentation." Applied Sciences, vol.9, no.11, pp. 2286 - 2300, June 2019.

[20] R. Caruana, A. Niculescu-Mizil, and G. Crew, "Ensemble selection from libraries of models." Proceedings of the twenty-first international conference on Machine learning. ACM, pp. 18, July 2004.

[21] Y. Liu, X. Chen, and C. Liu, "Delving into transferable adversarial examples and black-box attacks." arXiv preprint arXiv:1611.02770, 2016.

〈저자소개〉

민 정 기 (Jungki Min) 정회원
2018년 2월: 고려대학교 컴퓨터학과 졸업
2018년 3월~현재: 고려대학교 정보보호대학원 정보보호학과 석사과정
〈관심분야〉 정보보호, 시스템 보안, 기계학습

문 종 섭 (Jong-sub Moon) 종신회원
1981년 2월: 서울대학교 계산통계학과 학사
1983년 2월: 서울대학교 계산통계학과 석사
1991년 2월: Illinois Institute of Technology 전산학과 박사
1993년 3월~현재: 고려대학교 전자 및 정보공학부 교수
2001년 2월~현재: 고려대학교 정보보호대학원 겸임교수
〈관심분야〉 정보보호, 운영체제, 침입탐지