Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.31, NO.6, Dec. 2021 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2021.31.6.1215

선형 판별 분석 및 k-means 알고리즘을 이용한 적대적 공격 유형 분류 방안*

최 석 환,[1†] 김 형 건,[1] 최 윤 호[2‡]
[1,2]부산대학교 (대학원생, 교수)

An Adversarial Attack Type Classification Method Using Linear Discriminant Analysis and k-means Algorithm*

Seok-Hwan Choi,[1†] Hyeong-Geon Kim,[1] Yoon-Ho Choi[2‡]
[1,2]Pusan National University (Graduate student, Professor)

요 약

인공지능 기술은 우수한 성능을 기반으로 다양한 분야에 적용되고 있지만 입력 데이터에 인간이 감지할 수 없는 적대적 섭동을 추가하여 인공지능 모델의 오작동을 유도하는 적대적 예제에 취약하다. 현재까지 적대적 예제에 대응하기 위한 방법은 세 가지 범주로 분류할 수 있다. (1) 모델 재학습 방법; (2) 입력 변환 방법; (3) 적대적 예제 탐지 방법. 이러한 적대적 예제에 대응하기 위한 방법은 끊임없이 등장하고 있지만 각 적대적 공격 유형을 분류하는 연구는 미비한 실정이다. 따라서, 본 논문에서는 차원 축소와 군집화 알고리즘을 활용한 적대적 공격 유형 분류 방법을 제안한다. 구체적으로, 제안하는 방법은 적대적 예시로부터 적대적 섭동을 추출하고 선형 판별 분석(LDA)를 통해 적대적 섭동의 차원을 축소한 후에 k-means 알고리즘으로 적대적 공격 유형 분류를 수행한다. MNIST 데이터셋과 CIFAR-10 데이터셋을 대상으로 한 실험을 통해, 제안하는 기법은 5개의 적대적 공격(FGSM, BIM, PGD, DeepFool, C&W)을 효율적으로 분류할 수 있으며, 적대적 예제에 대한 정상 입력을 알 수 없는 제한적인 상황에서도 우수한 분류 성능을 나타내는 것을 확인하였다.

ABSTRACT

Although Artificial Intelligence (AI) techniques have shown impressive performance in various fields, they are vulnerable to adversarial examples which induce misclassification by adding human-imperceptible perturbations to the input. Previous studies to defend the adversarial examples can be classified into three categories: (1) model retraining methods; (2) input transformation methods; and (3) adversarial examples detection methods. However, even though the defense methods against adversarial examples have constantly been proposed, there is no research to classify the type of adversarial attack. In this paper, we proposed an adversarial attack family classification method based on dimensionality reduction and clustering. Specifically, after extracting adversarial perturbation from adversarial example, we performed Linear Discriminant Analysis (LDA) to reduce the dimensionality of adversarial perturbation and performed K-means algorithm to classify the type of adversarial attack family. From the experimental results using MNIST dataset and CIFAR-10 dataset, we show that the proposed method can efficiently classify five tyeps of adversarial attack(FGSM, BIM, PGD, DeepFool, C&W). We also show that the proposed method provides good classification performance even in a situation where the legitimate input to the adversarial example is unknown.

Keywords: Deep Learning, Adversarial example, Adversarial attack, Clustering

Received(10. 22. 2021), Modified(11. 19. 2021), Accepted(11. 23. 2021)

* 본 연구는 4단계 BK21사업 동남권4차산업혁명리더양성사업단과 부산대학교 기본연구지원사업(2년)에 의해 지원되었습니다.

† 주저자, dainailsh@pusan.ac.kr
‡ 교신저자, yhchoi@pusan.ac.kr(Corresponding author)

I. 서 론

인공지능 분야는 최근 몇 년 동안 산업, 교육, 경제 등 다양한 분야에서 우수한 성능을 보이며 널리 사용되고 있다. 최근 발표된 글로벌 시장조사기관 Tractica의 분석 보고서[1]에는 세계 인공지능 시장 규모가 2018년 95억 달러를 달성했으며 이후 약 43.4% 씩 증가하여 2025년에는 1,186억 달러 규모를 달성할 것으로 전망한다고 기술되어 있다. 이러한 분석 결과는 현 인공지능 분야의 발전을 잘 보여주는 사례이다.

하지만, 최근 인공지능 분야에서는 입력 이미지에 특정 노이즈를 추가하여 인공지능 모델의 분류 정확도를 크게 감소시키는 적대적 예제(Adversarial Examples)와 이를 수행하는 일련의 과정인 적대적 공격(Adversarial Attack)이 큰 이슈가 되고 있다[2]. 2018년 IBM의 적대적 공격 관련 toolbox 공개[3] 및 구글 브레인의 “앞으로 인공지능 모델을 어떻게 방어할 것인가가 큰 과제가 될 것”이라 직접 언급[4]하는 등 적대적 예제 및 이와 관련한 연구가 필요함을 시사한다.

적대적 예제가 이슈가 된 이래로 다양한 적대적 공격 기법[5][6][7][8][9]과 이에 대응하기 위한 방법이 꾸준히 등장하고 있다. 현재까지 적대적 공격에 대응하기 위한 방법은 크게 모델 재학습 방법[10][11]과 입력 변환 방법[12][13], 적대적 예제 탐지 방법[14][15][16] 세 가지로 분류할 수 있다. 모델 재학습 방법은 인공지능 모델의 견고성(robustness) 향상을 위해 인공지능 모델을 재학습 또는 새로운 인공지능 모델을 학습하는 방법이다. 이러한 모델 재학습 방법은 인공지능 모델 학습 과정에 포함된 적대적 예제에 대해 좋은 성능을 나타내지만, 학습 과정에 포함되지 않은 적대적 예제에 대해 낮은 방어 성능을 보인다. 즉, 모델 재학습 방법은 변종 적대적 공격 기법들에 대한 대응이 어렵다. 입력 변환 방법은 인공지능 모델에 대한 적대적 예제의 영향력을 감소하기 위해 입력 데이터를 변환하는 방법이다. 입력 변환 방법은 인공지능 모델의 재학습 없이 적대적 예제에 대한 대응이 가능하지만 적대적 예제뿐만 아니라 정상 입력에 대해서도 입력 변환을 수행하기 때문에 정상 입력에 대한 성능 감소가 발생하는 문제점이 있다[17]. 적대적 예제 탐지 방법은 인공지능 모델의 입력이 정상 입력인지 적대적 예제인지 판단함으로써 적대적 예제에 대응하는 방법이다. 이러한 적대적 예제 탐지 방법은 높은 방어 성능을 제공하지만, 탐지 결과가 단순한 이진 분류 (Binary Classification) 형태로만 제공되는 단점이 있다.

지금까지 적대적 예제에 대응하기 위한 방법은 끊임없이 등장하고 있지만 각 적대적 공격 유형을 분류하는 연구는 미비한 실정이다. 보안 분야에서 취약점 및 공격 유형에 대한 분류는 공격 완화 방안 설계, 피해에 대한 평가, 기존 방어 기법의 강화 등 많은 측면에서 매우 중요하다[18]. 예를 들어, 취약점 및 공격 유형 분류 결과를 기반으로 기존 대응 전략을 개선하거나 각 공격 유형에 적합한 대응 전략을 선택적으로 적용하여 대응 전략의 성능을 향상시킬 수 있다[19]. 또한, 취약점 및 공격 유형 분류는 새로운 취약점 및 공격에 대한 빠른 대응이 가능하도록 한다. 예를 들어, 새롭게 발견된 공격이 기존 공격 유형에 속한다면, 해당 공격 유형에서 자주 발생한 변종에 대한 사전 대응이 가능하다.

이미 많은 분야에서 이러한 취약점 및 공격 유형분류에 대한 중요성을 토대로 연구가 진행되고 있지만[20][21], 현재까지 적대적 공격에 대한 유형 분류 연구는 진행되지 않고 있다. 따라서, 본 논문에서는 차원 축소 기법과 군집 알고리즘을 활용하여 적대적 공격 유형을 효율적으로 분류하는 방법을 제안한다.

본 논문의 구성을 요약하면 다음과 같다. 2장에서는 대표적인 적대적 공격 기법과 적대적 공격에 대한 대응 기법에 대해 소개하고, 3장에서는 제안하는 적대적 공격 유형 분류 기법에 대해 상세히 기술한다. 4장에서는 제안하는 기법에 대한 성능 검증 결과를 기술하고, 5장에서는 전체적인 내용을 요약 기술한다.

II. 관련 연구

2.1 적대적 공격 기법

Ian Goodfellow 등은 경사 하강법(Gradient Descent)을 역방향으로 수행하여 적대적 예제를 생성하는 Fast Gradient Sign Method(FGSM)을 제안하였다[5]. FGSM은 빠르게 적대적 예제를 생성할 수 있지만 공격 성공률이 낮은 문제점이 존재하였다. 이를 해결하기 위해 Kurakin 등은 FGSM을 반복적으로 수행하는 Basic Iterative Method(BIM)을 제안하였고[6], Madry 등은 BIM을 일반화한 Projected Gradient Descent(PGD)를 제안하였다[7]. BIM과 PGD

모두 높은 공격 성공률을 보였으나 적대적 예제 생성에 더해지는 적대적 섭동(Adversarial Perturbation)의 크기가 크다는 한계점이 존재하였다. 이러한 적대적 섭동의 크기를 줄이기 위해, Moosavi-Dezfooli 등은 공격 대상 모델의 결정 경계와 가장 근접한 적대적 예제를 생성하는 DeepFool을 제안하였다[8]. 또한, Carlini와 Wagner 등은 다수의 섭동 후보를 계산하고 크기가 가장 작은 섭동을 최종적으로 선택하는 C&W를 제안하였다[9]. C&W는 distance metric(l_∞, l_0, l_2)에 따라 3가지 공격으로 나뉠 수 있으며, 본 논문에서는 많은 논문의 성능 검증에 사용된 l_2 기반의 C&W를 성능 검증에 고려하였다[13][17].

2.2 적대적 공격 대응 기법

2.2.1 모델 재학습 방법

모델 재학습 방법은 적대적 공격에 대응하기 위해 인공지능 모델을 재학습하거나 새로운 인공지능 모델을 학습하는 방법이다. 대표적인 모델 재학습 방법으로는 Adversarial Training[10]과 Defensive Distillation[11]이 있다.

Adversarial Training은 인공지능 모델의 견고성(robustness)을 개선시키기 위해 기존의 학습 데이터셋에 적대적 예제를 추가하여 인공지능 모델을 재학습시키는 대표적인 방어 방법이다. Defensive Distillation은 기존 모델을 재학습 하지 않고, 두 개의 모델(기존 모델, 기존 모델을 모방한 새로운 모델)을 학습하여 적대적 예제에 대해 견고성을 향상시킬 수 있도록 인공지능 모델의 구조를 변경하는 방법이다. 상기 언급한 두 방법은 구현이 간단하지만 현재 배포된 인공지능 모델을 재학습하는 데 많은 비용이 필요한 문제가 있다.

2.2.2 입력 변환 방법

입력 변환 방법은 Filtering 및 Denoising과 같은 이미지 변환 기법을 적용하여 적대적 예제의 영향력을 감소시키는 방법이다. 대표적인 입력 변환 방법으로는 MagNet[12]과 PixelDefend[13]가 있다.

MagNet은 적대적 예제의 섭동을 줄이기 위해 오터인코더 기반의 reformer를 사용하여 적대적 예제를 정상 입력의 분포로 매핑하는 방법이다. PixelDefend는 Pixel-CNN 모델 기반의 이미지 Denoising 모듈을 통해 적대적 예제의 섭동 크기를 줄이는 방법이다. 상기 언급한 두 방법 모두 기존의 인공지능 모델을 재학습 또는 변경하지 않고 그대로 사용 가능하다. 하지만 두 방법 모두 정상 입력에 대해서도 이미지 변환 기법을 적용하기 때문에 정상 입력에 대한 인공지능 모델의 성능 감소가 발생하는 문제가 있다.

2.2.3 적대적 예제 탐지 방법

적대적 예제 탐지 방법은 인공지능 모델의 입력이 적대적 예제일 확률을 계산하거나 적대적 예제와 정상 입력 간의 분포를 비교하여 적대적 공격이 발생했는지 여부를 탐지하는 방법이다.

Jiajun Lu등 은 기존 인공지능 모델에 적대적 예제 탐지를 위한 별도의 인공지능 모델을 연결한 SafetyNet을 제안하였다[14]. SafetyNet의 탐지 모델은 인공지능 모델의 마지막 활성화 함수의 출력을 양자화(Quantization) 하고 양자화 된 결과를 기반으로 적대적 예제와 정상적인 입력의 분포 차이를 학습한다. Carrara 등은 적대적 예제와 정상 입력은 인공지능 모델 내에서 활성화되는 위치가 다르다는 점에 기반하여 인공지능 모델 내 각 계층에서 추출한 중간 특징들을 활용해 적대적 예제의 발생 여부를 탐지하는 방법을 제안하였다[15]. Xu 등은 인공지능 모델의 입력 데이터를 적대적 예제 또는 정상적인 입력 중 하나로 이진 분류하기 위한 Feature Squeezing 방법을 제안하였다[16]. Feature Squeezing은 Squeezer와 Detector로 구성되어 있으며 Squeezer는 입력 데이터를 Squeezed data로 변환하여 적대적 예제의 섭동 크기를 줄이고 Detector는 이러한 Squeezed data를 사용하여 적대적 공격 발생 여부를 탐지한다. Zheng 등은 군집화 알고리즘을 사용하여 인공지능 모델의 입력 데이터가 적대적 예제인지 정상적인 입력인지 이진 분류하는 방안을 제안하였다[28].

상기 언급한 적대적 예제 탐지 방법들은 모두 적대적 공격의 발생 여부만 탐지가 가능하며 발생한 적대적 공격의 유형에 대해서는 분류가 불가능하다. 따라서, 본 논문에서는 적대적 공격 유형에 대한 분류가 가능한 방법을 제안한다.

III. 제안 모델

이 장에서는 제안하는 적대적 공격 유형 분류 방법에 대해 기술한다. 제안하는 방법은 크게 세 절차를 통해 수행된다: (1) 적대적 섭동 추출 단계; (2) 차원 축소 단계; (3) 군집화 단계.

3.1 적대적 섭동 추출 단계

적대적 섭동 추출 단계에서는 정상 입력과 적대적 예제의 산술 뺄셈 연산을 통해 적대적 공격에 의해 추가된 적대적 섭동을 추출한다.

Fig. 1은 MNIST 데이터셋[22]에서 하나의 테스트 샘플에 대한 5개의 적대적 공격(FGSM, BIM, PGD, DeepFool, C&W)의 결과와 각 적대적 공격으로부터 추출된 적대적 섭동을 나타낸다. Fig. 2는 CIFAR-10 데이터셋[23]에서 하나의 테스트 샘플에 대한 5개의 적대적 공격의 결과와 각 적대적 공격으로부터 추출된 적대적 섭동을 나타낸다.

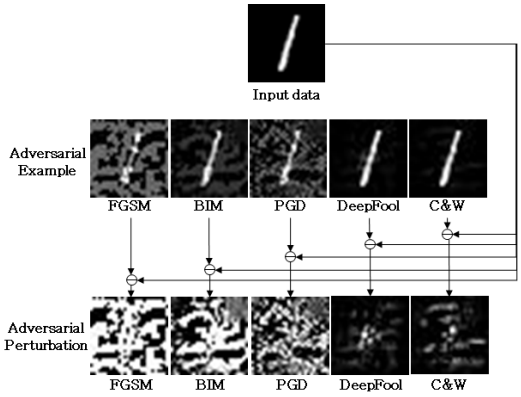


Fig. 1. Some examples of adversarial examples and extracted adversarial perturbations using MNIST test dataset

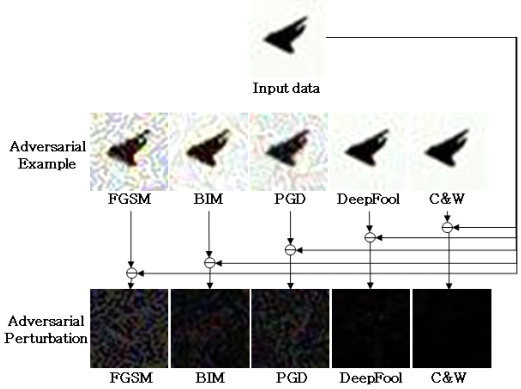


Fig. 2. Some examples of adversarial examples and extracted adversarial perturbations using CIFAR-10 test dataset

3.2 차원 축소 단계

적대적 섭동 추출 단계에서 추출된 섭동을 직접 사용하여 적대적 공격 유형을 분류하는 것은 추출된 적대적 섭동이 불필요한 값을 포함하기 때문에 비효율적이다. 따라서, 차원 감소 단계에서는 추출된 적대적 섭동으로부터 주요 특징을 추출하기 위해 적대적 섭동의 차원을 축소한다. 구체적으로, 제안하는 모델은 차원 축소를 위해 선형 판별 분석(LDA)[24]을 수행한다.

선형 판별 분석은 같은 유형 내 샘플 간의 분산을 최소화하면서 다른 유형 내 샘플 간의 분산을 최대화하는 지도학습 기반의 차원 축소 방법이다. 따라서, 선형 판별 분석은 서로 다른 적대적 공격 유형을 분리하는 방향으로 투영(projection)시키는 것에 적합하다. Fig. 3은 제안하는 모델에서 사용하는 선형

Algorithm 1 LDA(Linear Discriminant Analysis)

Input: X[N][M] - Set of N data items,
N – Number of samples,
M – Number of dimensions in the sample
C – Number of class in the samples

Output: XP - Set of projected X

```
procedure LinearDscriminantAnlaysis(X):
    Initialize M[M], MC[C][M], SB[M][M], SW[C][M][M], λ[C], V[C][M][M],
    M = Means(X)            // Mean for all data
    MC = ClassMeans(X)      // Mean for each class
    for i in C: // Calculate between-class matrix loop
        SB += N_i(MC[i] – M)(MC[i] – M)^T
    for i in C:  // Compute within-class matrix loop
        SW[i] ← (X_i – MC[i])(X_i – MC[i])^T
    for i in C:
        for j in C:
            ω ← (SW[i])^-1 SB // Transformation Matrix
            λ[i],V[i]←Eigen(ω) //Eigenvalue, Eigenvector
    V_OPT = Optimization(λ, V)
    XP = V_OPT^T X
end procedure
return XP
```

Fig. 3. LDA(Linear Discriminant Analysis) Algorithm

판별 분석 알고리즘의 동작 절차를 나타낸다. 추출된 적대적 섭동의 집합이 주어졌을 때, 추출된 적대적 섭동 전체 평균값 M과 적대적 공격 유형별 평균값 MC를 구한다(line 1-4). 이후 적대적 공격 유형 간의 분산 값으로 산포행렬 SB를 구성하고(line 5-6) 동일한 유형의 내 샘플들 간의 분산 값으로 산포행렬 SW를 구성한다(line 7-8). 구성된 산포행렬 SW와 SB을 사용하여 변환 행렬인 ω와 고유값 λ와 고유벡터 V를 계산하고(line 9-12) λ가 최대가 되는 V를 탐색한다(line 13). 이때, 적대적 공격 유형 간 분산 SW가 가장 크고, 동일 유형 샘플 간 분산 SB가 최소가 될 때, λ의 값이 최대가 된다. 마지막으로, 추출된 적대적 섭동의 집합은 생성된 고유벡터 V_{OPT} 를 사용하여 새로운 차원으로 투영된다(line 14-16).

3.3 군집화 단계

군집화 단계에서는 저차원의 단순화된 적대적 섭동의 주요 특징과 군집화 알고리즘을 사용하여 적대적 공격 유형 분류를 수행한다. 적대적 공격 유형 분류를 위해 군집화 알고리즘을 사용한 이유는 다음과 같다.

첫째, 비지도 학습인 군집화 알고리즘의 특징을 이용하여 적대적 공격을 다양한 유형으로 분류할 수 있다. 예를 들어, 군집의 수를 낮게 설정하면 각 공격 유형을 큰 범위에서 분류할 수 있고, 군집의 수를 높게 설정하면 각 공격 유형을 세부적으로 분류할 수 있다. 이는 차원 축소 방법과 결합하여 다양한 관점에서의 적대적 공격 유형의 분석을 가능하게 한다.

둘째, 새로운 공격에 대한 대응을 가능하게 한다. 새로운 공격 유형에 대한 분류가 불가능한 지도학습 기반의 분류 알고리즘과 달리, 군집화 알고리즘은 파라미터 조정을 통해 새로운 적대적 공격 유형을 분류할 수 있다.

본 논문에서는 k-means 알고리즘[25]을 사용하여 군집화를 수행한다. k-means 알고리즘은 다양한 응용 분야에서 사용되는 가장 단순한 군집화 방법 중 하나이다. 입력 데이터에 대한 군집을 찾기 위해, k-means 알고리즘은 전체 데이터 세트에서 서로 유사한 데이터를 군집화하고 분류한다. Fig. 4는 제안하는 모델에서 사용하는 k-means 알고리즘의 군집화 절차를 나타낸다. 저차원의 단순화된 적대적 섭동들로부터 임의로 선택된 K개의 항목을 초기 중심

Algorithm 2 k-Means clustering algorithm

Input: D - Set of N data items
K - Number of clusters to form

Output: C : Set of K clusters

```
1  procedure kMeansClusteringAlgorithm(D, K):
2      Arbitrarily choose K data items from D as initial centroids;
3      do
4          Assign each item of D_i to the clusters which has the closet centroid;
5          Calculate new mean for each cluster;
6      while Convergence criteria is met;
7  end procedure
8  return C
```

Fig. 4. k-means Algorithm

점으로 지정한다(line 1-2). 이 후, 각 항목을 가장 가까운 중심점의 군집에 할당한다(line 3-4). 각 군집 내의 항목 평균을 계산하여 군집 중심을 갱신한다(line 5). 수렴 기준으로 각 항목들이 할당된 소속 군집의 변경 여부를 확인하고, 수렴하지 않으면 각 항목의 군집 할당 및 군집 중심 갱신을 수렴할 때까지 반복한다(line 6).

IV. 실험 결과

본 장에서는 MNIST 데이터셋과 CIFAR-10 데이터셋을 이용하여 제안한 방법의 성능을 검증한다.

4.1 실험 환경

본 논문에서는 제안하는 방법을 Python v3.6.9, Tensorflow v1.14, Keras v2.3.1, Cleverhans v2.1.0[26]을 활용하여 구현 및 실험하였다. 제안하는 방법의 학습 및 성능 평가를 수행한 컴퓨팅 환경은 Table 1과 같다.

본 논문에서 사용하는 인공지능 모델은 다음과 같

Table 1. Experimental environments

	Environment
OS	Ubuntu 18.04.3 LTS
Kernel	5.3.0-62-generic
CPU	2.40GHz CPU clock (Intel(R) Xeon(R) CPU E5-2630 v3)
GPU	GeForce RTX 2080 Ti
RAM	64GB

다. MNIST 데이터셋의 경우, 회색조(Gray-scale) 영상으로 24비트의 RGB 컬러 형식에 비해 8비트의 비교적 단조로운 형식을 가지고 있기 때문에 적대적 공격 대상 심층신경망은 간단한 구조의 CNN(Convolutional Neural Network) 모델을 사용하였다. 구체적으로, 실험에 사용한 CNN 모델은 5x5 크기의 커널 및 32개의 필터를 갖는 합성곱(Convolutional) 계층 그리고 ReLU 활성화 함수, 2x2 크기의 Max Pooling 계층, Flatten 계층, 64개의 필터를 갖는 완전 연결(Fully-connected) 계층 및 ReLU 활성화 함수, 10개의 필터를 갖는 완전 연결 계층 및 소프트맥스(Softmax) 함수로 구성된다. 해당 모델의 학습 정확도는 0.995, 검증 정확도는 0.980, 시험 정확도는 0.981을 보였다. 반면, CIFAR-10 데이터셋의 경우, 24비트의 RGB 컬러 형식으로 MNIST보다 비교적 복잡한 형식을 가지고 있다. 따라서, MNIST 데이터셋 실험에 사용된 적대적 공격 대상 심층신경망보다 계층이 깊고 복잡한 표준 심층신경망 중 하나인 ResNet-34[27]로 사용하였다. 해당 모델의 학습 정확도는 0.783, 검증 정확도는 0.761, 시험 정확도는 0.756을 보였다.

또한, 본 논문의 실험에서는 Cleverhans 라이브러리를 이용해 많은 관련 논문에서 활용되는 벤치마크 공격 기법인 FGSM[5], BIM[6], PGD[7], DeepFool[8], C&W[9]에 대한 분류를 수행하였으며, 각 적대적 공격 유형 당 1,000개의 샘플을 생성하였다. 각 적대적 공격에 대한 매개변수는 대부분의 적대적 공격 대응 기법[16][17] 및 Cleverhans에서 기본으로 제공하는 매개변수에 적대적 공격 성능 향상을 위해 적대적 섭동의 크기 및 반복횟수를 조정하여 사용했으며 이를 Table 2에서 기술하였다.

Table 2. Adversarial attack parameters

Adversarial attack	parameter
FGSM	epsilon=0.2, clip_min=0.0, clip_max=1.0
BIM	epsilon_iter=0.01, nb_iter=20, clip_min=0.0, clip_max=1.0
PGD	epsilon_iter=0.01, nb_iter=20, clip_min=0.0, clip_max=1.0
DeepFool	clip_min=0.0, clip_max=1.0, overshoot=0.2, max_iter=80
C&W	clip_min=0.0, clip_max=1.0, binary_search_steps=1, initial_const=0.06, max_iterations=300, learning_rate=0.2

4.2 성능 평가 지표

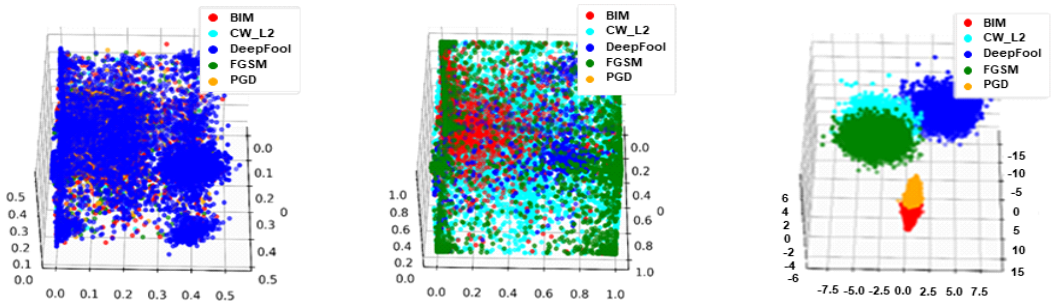
제안하는 기법의 성능을 평가하기 위해, 본 논문에서는 대표적인 군집 알고리즘 성능 평가 지표인 Homogeneity(동질성), Completeness(완전성), V-measure를 측정하였다. Homogeneity는 각 군집이 각 적대적 공격 유형의 데이터(data points)만을 포함하는 정도이다. Completeness는 각 적대적 공격 유형의 모든 데이터 점들이 동일한 군집 내에 있는 정도이다, V-measure는 동질성과 완전성 점수 사이의 조화 평균(Harmonic mean)을 계산한 값이다. 세 지표 모두 0.0 ~ 1.0 사이의 값으로 나타낼 수 있으며, 그 값이 클수록 성능이 더 좋음을 의미한다.

4.3 실험 결과

현재까지 적대적 공격 유형을 분류하는 연구 사례가 없기 때문에 제안하는 기법의 성능 평가를 위한 비교 대상이 존재하지 않는다. 따라서, 본 논문에서는 제안하는 기법의 성능을 검증하기 위해 다음과 같은 자체적인 비교 대상을 정의하고 성능을 비교하였다: (1) 적대적 예제 자체에 k-means 알고리즘을 적용한 경우(Step 3); (2) 적대적 섭동을 추출해 k-means 알고리즘을 적용한 경우(Step 1 + Step 3).

Fig. 5은 MNIST 데이터셋에 대해 제안하는 기법과 비교 대상과의 적대적 공격 유형 분류 성능을 시각화하여 나타낸 결과이다. 적대적 예제 자체에 k-means 알고리즘을 적용한 경우(Fig. 5(a))와 적대적 섭동을 추출해 k-means 알고리즘을 적용한 경우(Fig. 5(b))는 적대적 공격의 유형을 분류하지 못하는 것을 확인할 수 있으며 제안하는 기법(Fig. 5(c))은 적대적 공격의 유형을 분류하는 것을 확인할 수 있다.

Fig. 6는 CIFAR-10 데이터셋에 대한 시각화 결과이다. MNIST 데이터셋에 대한 결과와 마찬가지로, 적대적 예제 자체에 k-means 알고리즘을 적

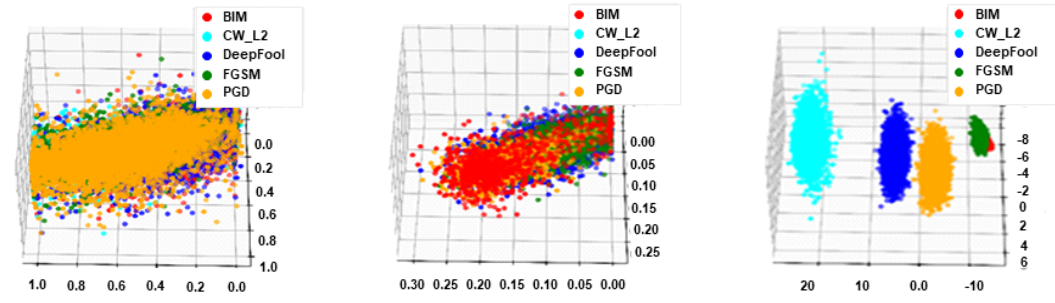


(a) adversarial example +k-means (Step 3)

(b) adversarial perturbation +k-means (Step 1 + Step 3)

(c) adversarial perturbation +LDA + k-means (Step 1 + Step 2 +Step3)

Fig. 5. Comparison of clustering visualization results using MNIST dataset



(a) adversarial example +k-means (Step 3)

(b) adversarial perturbation +k-means (Step 1 + Step 3)

(c) adversarial perturbation +LDA + k-means (Step 1 + Step 2 +Step3)

Fig. 6. Comparison of clustering visualization results using CIFAR-10 dataset

용한 경우(Fig. 6(a))와 적대적 섭동을 추출해 k-means 알고리즘을 적용한 경우(Fig. 6(b))는 적대적 공격의 유형을 분류하지 못하는 반면, 제안하는 기법(Fig. 6(c))은 CIFAR-10 데이터셋에 대해서도 적대적 공격의 유형을 분류하는 것을 확인할 수 있다.

그림 Fig. 5(c)와 Fig. 6(c)에서 BIM과 PGD는 인접한 좌표상에 군집이 형성되어있는 것을 확인할 수 있다. 이는 PGD가 BIM을 일반화한 적대적 공격이므로 두 적대적 공격 기법이 유사한 특성을 공유하기 때문이다. 이러한 결과를 통해 제안하는 기법이 적대적 공격의 특성을 반영한 유형 분류를 수행하는 것을 알 수 있다.

Table 3은 각 데이터셋에 대해 제안하는 기법과

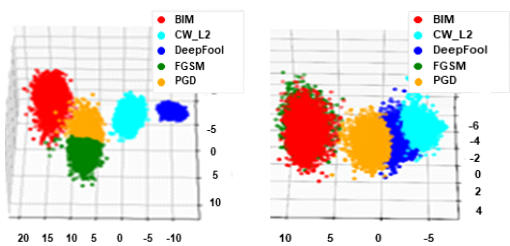
Table 3. Comparison results using clustering evaluation metrics

	metric	MNIST dataset	CIFAR-10 dataset
adversarial example + k-means (Step 3)	Homogeneity	0.247	0.000
	Completeness	0.252	0.000
	V-measure	0.249	0.000
adversarial perturbation +k-means (Step 1 + Step 3)	Homogeneity	0.601	0.823
	Completeness	0.654	0.927
	V-measure	0.626	0.872
adversarial perturbation +LDA + k-means (Step 1 + Step 2 + Step3)	Homogeneity	0.873	0.925
	Completeness	0.874	0.926
	V-measure	0.874	0.925

비교 대상의 Homogeneity, Completeness, V-measure를 측정한 결과이다. 적대적 예제 자체에 k-means 알고리즘을 적용한 경우(Step 3)에는 군집화 성능이 매우 낮은 것을 확인할 수 있다. 특히, CIFAR-10 데이터셋에 대해서는 군집화 형성이 제대로 이루어지지 않았음을 실험 결과를 통해 알 수 있다. 적대적 섭동을 추출해 k-means 알고리즘을 적용한 경우(Step 1 +Step 3)에는 적대적 예제 자체에 k-means 알고리즘을 적용한 경우보다 좋은 군집화 성능을 보였지만 제안하는 기법에 비해 현저히 낮은 군집화 성능을 보였다. 제안하는 기법의 경우에는 두 데이터셋 모두에 대해 효율적인 군집화 성능을 보이는 것을 확인할 수 있다.

제안하는 방법은 적대적 섭동을 추출하기 위해 정상 입력과 적대적 예제의 산술 뺄셈 연산을 수행한다. 그러나, 인공지능 모델의 실제 적용 환경에서는 입력 데이터에 적대적 공격이 발생했는지 또는 어떤 공격이 수행되었는지 알 수 없기 때문에 적대적 예제에서 직접 적대적 섭동을 얻는 것은 불가능하다. 이러한 상황에서, 제안하는 기법은 Binary filter, Median smoothing filter와 같은 잡음 제거 기법을 사용하여 적대적 섭동을 얻을 수 있다. 구체적으로, 적대적 섭동을 추출하기 위해 제안하는 기법은 잡음 제거 기법이 적용된 적대적 예제와 적대적 예제 간 산술 뺄셈 연산을 수행함으로써 적대적 섭동을 얻을 수 있다.

Fig. 7는 이러한 제한적인 상황에서 잡음 제거 기법을 적용한 제안하는 기법의 적대적 공격 유형 분류 성능을 나타낸 결과이다. MNIST 데이터셋에 경우에는 Binary Filter를, CIFAR-10 데이터셋의 경우에는 Median smoothing filter를 적용하였다. 실험 결과에서 확인할 수 있듯이, 제한적인 상황에서도 제안하는 기법은 적대적 공격 유형을 효율적으로 분류할 수 있다.



Homogeneity : 0.827
Completeness : 0.902
V-measure : 0.863
(a) MNIST 데이터셋
(Binary Filter)

Homogeneity : 0.678
Completeness : 0.729
V-measure : 0.703
(b) CIFAR-10 데이터셋
(Median Filter)

Fig. 7. Clustering performance of the proposed method under the restricted scenario

V. 결 론

본 논문에서는 적대적 예제에 대한 대응 방법들의 한계를 파악하고 이를 개선하기 위한 적대적 공격 유형 분류 방법을 제안하였다. 제안하는 적대적 공격 유형 분류 방법은 선형 판별 분석(LDA)에 기반한 차원 축소 및 k-means 알고리즘에 기반한 군집화를 통해 다양한 유형의 적대적 공격을 효율적으로 분류할 수 있으며, 이를 MNIST 데이터셋과 CIFAR-10 데이터셋을 활용한 실험을 통해 검증하였다. 또한, 제안하는 기법은 적대적 예제에 대한 정상 입력을 알 수 없는 환경에서도 적대적 공격 유형 분류가 가능한 장점을 가진다.

그러나 본 논문에서 제안한 적대적 공격 유형 분류 방법의 군집화에서 사용되는 k-means 알고리즘은 사전에 공격 유형의 수를 정해야하는 한계점이 존재한다. 따라서 향후 연구에서는 DBSCAN 알고리즘과 같이 공격 유형의 수를 정할 필요가 없는 군집화 알고리즘을 사용하여 새로운 공격 유형에도 유연하게 대응 가능하도록 제안하는 방법을 확장할 것이다. 또한, 단순 적대적 공격 유형 분류를 넘어 적대적 공격 탐지 와 유형 분류를 동시에 할 수 있는 대응 방안에 대한 연구도 수행할 것이다.

References

[1] Tractica, "Artificial Intelligence Market Forecasts," Dec. 2019.

[2] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, Dec. 2013.

[3] Nicolae, M. I., Sinn, M., Minh, T. N., Rawat, A., Wistuba, M., Zantedeschi, V., & Edwards, B. "Adversarial Robustness Toolbox v0. 2.2.," Jul. 2018.

[4] Jackie SnowMar. "To protect artificial

intelligence from attacks, show it fake data," Mar. 2018.
〔5〕 Goodfellow, I. J., Shlens, J., & Szegedy, C.. "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, Dec. 2014
〔6〕 Kurakin, A., Goodfellow, I., & Bengio, S. "Adversarial examples in the physical world," Jul. 2016.
〔7〕 Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, Jun. 2017.
〔8〕 Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. "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, 2016
〔9〕 Carlini, N., & Wagner, D. "Towards evaluating the robustness of neural networks," In 2017 ieee symposium on security and privacy (sp), pp. 39-57, May. 2017
〔10〕 Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, Jun. 2017.
〔11〕 Carlini, N., & Wagner, D. "Defensive distillation is not robust to adversarial examples," arXiv preprint arXiv:1607.04311, Jul. 2016.
〔12〕 Guo, C., Rana, M., Cisse, M., & Van Der Maaten, L. "Countering adversarial images using input transformations," arXiv preprint arXiv:1711.00117, Oct. 2017.
〔13〕 Song, Y., Kim, T., Nowozin, S., Ermon, S., & Kushman, N. "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," arXiv preprint arXiv:1710.10766, Oct. 2017.
〔14〕 Lu, J., Issaranon, T., & Forsyth, D. "Safetynet: Detecting and rejecting adversarial examples robustly," In Proceedings of the IEEE International Conference on Computer Vision, pp. 446-454, Aug. 2017.
〔15〕 Carrara, F., Becarelli, R., Caldelli, R., Falchi, F., & Amato, G. "Adversarial examples detection in features distance spaces," In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Sep. 2018.
〔16〕 Xu, W., Evans, D., & Qi, Y. "Feature squeezing: Detecting adversarial examples in deep neural networks," arXiv preprint arXiv:1704.01155, Apr. 2017.
〔17〕 Choi, S. H., Shin, J., Liu, P., & Choi, Y. H. "EEJE: Two-Step Input Transformation for Robust DNN Against Adversarial Examples," IEEE Transactions on Network Science and Engineering, 8(2), pp. 908-920. Jul. 2020
〔18〕 Mohaisen, A., West, A. G., Mankin, A., & Alrawi, O. "Chatter: Classifying malware families using system event ordering," In 2014 IEEE Conference on Communications and Network Security, pp. 283-291. Oct. 2014.
〔19〕 AlAhmadi, B. A., & Martinovic, I. "MalClassifier: Malware family classification using network flow sequence behaviour," In 2018 APWG Symposium on Electronic Crime Research (eCrime), pp. 1-13, May. 2018.
〔20〕 Ahmadi, M., Ulyanov, D., Semenov, S., Trofimov, M., & Giacinto, G. " Novel feature extraction, selection

and fusion for effective malware family classification," In Proceedings of the sixth ACM conference on data and application security and privacy, pp. 183-194. Mar. 2016.

[21] Alswaina, F., & Elleithy, K. "Android malware family classification and analysis: Current status and future directions," Electronics, 9(6), 942. 2020.

[22] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[23] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (Canadian Institute for Advanced Research). 2009.

[24] Balakrishnama, S., & Ganapathiraju, A. "Linear discriminant analysis-a brief tutorial," Institute for Signal and information Processing, 18, pp. 1-8. 1998

[25] Hartigan, J. A., & Wong, M. A. "Algorithm AS 136: A k-means clustering algorithm," Journal of the royal statistical society. series c (applied statistics), 28(1), pp. 100-108. 1979

[26] Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., & McDaniel, P. "Technical report on the cleverhans v2. 1.0 adversarial examples library," arXiv preprint arXiv:1610.00768, Oct. 2016.

[27] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770-778, Sep. 2016.

[28] Zheng, Yanbin, et al. "Defence against adversarial attacks using clustering algorithm," International Conference of Pioneering Computer Scientists, Engineers and Educators. Springer, Singapore, Sep. 2019.

〈저자소개〉

최 석 환 (Seok-Hwan Choi) 학생회원
2016년: 부산대학교 정보컴퓨터공학부 학사
2016년 9월~현재: 부산대학교 전자전기컴퓨터공학과 석박사통합 과정
〈관심분야〉 AI 보안, 침입탐지

김 형 건 (Hyeong-Geon Kim) 학생회원
2017년: 경성대학교 컴퓨터공학부 학사
2019년: 부산대학교 전기전자컴퓨터공학과 석사
2019년 9월~현재: 부산대학교 정보융합공학과 박사과정
〈관심분야〉 개인정보보호, 랜섬웨어

최 윤 호 (Yoon-Ho Choi) 종신회원
2008년: 서울대학교 전기컴퓨터공학부 박사
2010년: 펜실베니아 주립대학교 박사후 연구원
2012년: 삼성전자 네트워크사업부 책임연구원
2014년: 경기대학교 융합보안학과 조교수
2016년: 부산대학교 전기컴퓨터공학부 조교수
2016년~현재: 부산대학교 전기컴퓨터공학부 부교수
〈관심분야〉 지능형 악성코드분석, 코드보안, 소프트웨어 취약성분석, 프라이버시, AI보안, 블록체인, 유무선 네트워크 보안 등