# 신경망의 적대적 견고성을 위한 혼동 행렬 기반 비트 반전 알고리즘

장우람, 황우진, 박호성*

전남대학교 컴퓨터정보통신공학과, *ICT 융합시스템공학과

195985@jnu.ac.kr, 195910@jnu.ac.kr, *hpark1@jnu.ac.kr

# Confusion Matrix Based Bit Flipping Algorithm for Adversarial Robustness

Wooram Jang, Woojin Hwang, Hosung Park*

Chonnam National University

요 약

본 논문에서는 혼동 행렬을 사용한 코드북 최적화를 통해 적대적 견고성(adversarial robustness)을 향상시키는 방식을 제안하고자 한다. 채널 잡음과 적대적 예제(adversarial example)의 유사성을 근거로 오류 정정 출력 부호어(Error-Correcting output code: ECOC)를 적용한 적대적 학습 또한 연구된 바 있다. 적대적 견고성 향상을 위해 부호어 간 해밍 거리 최적화 또는 부호어 할당 알고리즘이 주로 사용된다. 본 연구에서는 클래스 간 구분 가능성을 고려한 부호어 비트 반전 알고리즘을 통해 선행 연구[4]에서 제시된 방식 대비 약 5%의 정확도 향상을 달성하였다.

## Ⅰ. 서 론

심층신경망(Deep Neural Network: DNN)은 다양한 분야에서 이미 훌륭한 성능이 입증되었다. DNN은 높은 성능을 보이는 동시에 의도적인 적대적 공격(adversarial attacks)에 대해 취약한 모습을 보인다. 따라서 DNN의 적대적 견고성을 확보하기 위한 적대적 학습, ECOC 등의 연구가 활발히 진행되고 있다.

하다마드 행렬은 ECOC에 주로 채택되어 왔다. 하지만 하다마드 행렬은 부호어 간 해밍 거리가 모두 같기 때문에, 클래스 간 구별하기 어려운 정도를 반영하기 힘들었다. 주요 선행 연구[1]에서는 두 가지 기법을 적용한 ECOC를 통해 DNN의 적대적 견고성을 향상시켰다. 해당 연구는 비트 반전을 통해 특정 부호어 간 해밍 거리를 증가시키며 코드북을 최적화한다. 이를 통해 추가적인 오류 정정 능력의 향상을 기대할 수 있다. 이후 오류 확률을 기준으로 구별하기 어려운 클래스에 높은 해밍 거리를 가지는 부호어를 할당하는 방식으로 문제를 해결한다. 이러한 방식은 적대적 공격에 대한 오류 정정 능력을 확보함으로써 적대적 견고성을 향상시켰다. 본 연구에서는 두 단계를 통합한 코드북 최적화 방식을 제안한다. 혼동 행렬을 활용함으로써 클래스 간 구분 가능성을 반영할 수 있으며, 모델 구조 단순화가 가능하다. 데이터 특성에 맞는 코드북 최적화를 통해 선행 연구 대비 약 5%의 정확도 향상을 도출하였다.

## Ⅱ. 본 론

### *2.1 배경*

DNN은 이미지 분류, 자연어 처리 등의 다양한 분야에서 이미 높은 성능을 입증한 바 있다. 하지만 적대적 공격에 대한 DNN의 취약점 또한 발견되었다. 적대적 예제란 인간은 알아차릴 수 없을 만큼 정교하게 제작된 변형이 원본 데이터에 추가된 것을 의미한다. 적대적 공격은 적대적 예제를 DNN 입력으로 사용하는 것을 의미하고 많은 경우 DNN에 심각한 오류를 초래한다. 적대적 견고성은 적대적 공격에 대해 방어할 수 있는 능력을 의미한다.

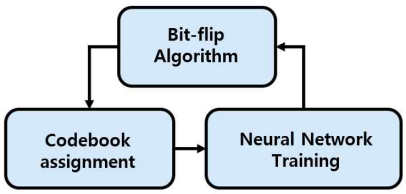


그림 1 ECOC 모델 구조

현재 적대적 공격을 검출하고 방어하는 연구가 활발히 진행되고 있으나 많은 경우 적대적 견고성은 정확도의 손실로부터 얻어지거나 다양한 적대적 공격에 대해 방어가 불가능하다는 한계에 봉착한다. 예를 들어 적대적 학습 기법은 높은 적대적 견고성을 얻을 수 있으나, 학습 과정에서 적대적 예제를 사용하기에 정확도 측면에서 성능 저하가 있다.

### *2.2 오류 정정 출력 부호어*

과거 다중 클래스 분류를 위한 앙상블 기법으로 distributed output code가 제안되었다[2]. 주로 신경망의 출력층에 채택되는 one-hot encoding과 softmax의 조합은 가장 높은 확률에 해당하는 클래스로 분류하는 방식이다. 반면 distributed output code는 각각의 클래스에 $n$길이의 부호어를 할당하고, 추론 결과와 부호어의 해밍 거리를 판단하는 방식이다. 복호기는 추론 결과와 가장 근접한 부호어를 가진 클래스로 분류한다. 이러한 방식은 이진 분류기의 앙상블 모델을 사용함으로써 기존의 방식에 비해 모델의 신뢰도를 높일 수 있었고, 클래스의 개수가 많아도 출력층의 뉴런을 줄일 수 있다는 점에서 효율적이다. 하지만 distributed output code는 부호어 간 최소 해밍 거리가 보장되지 않는 한계가 있었다.

ECOC의 적용은 적대적 예제의 변형과 채널 잡음의 유사성에 의해 고려되었다. ECOC는 오류 정정 부호를 부호어로 사용함으로써 최소 해밍 거리를 확보하고 오류 정정 능력을 보장한다. 기존 적대적 학습 방법의 문제점은 원본 데이터에 대한 정확도 손실과 매 반복마다 적대적 예제를 생성해 훈련해야 했다는 것이다. 반면 ECOC는 원본 데이터를 통해 훈련을 진행한다. 따라서 원본 데이터에 대한 정확도를 유지할 수 있으며, 훈련

시간 또한 획기적으로 줄일 수 있다. ECOC에 사용되는 오류 정정 부호에서 최소 해밍 거리를 $d_H$라고 할 때, 최대 $\lfloor \frac{d_H-1}{2} \rfloor$ 개의 오류 정정 능력이 보장된다. 따라서 적대적 공격으로 인해 $\lfloor \frac{d_H-1}{2} \rfloor$ 개 이하의 오류가 발생한 경우 정정이 가능하기에 정확도를 유지할 수 있다.

주요 선행 연구[1]에서는 적대적 견고성 향상을 위해 ECOC 모델의 구조를 두 단계로 나눈 후 최적화를 진행했다. 먼저 부호어 간 해밍 거리를 증가시키는 방식으로 코드북을 최적화한 후, 이를 적절한 클래스에 할당하는 방식을 사용하여 적대적 견고성을 향상시켰다. 본 연구에서는 혼동 행렬을 고려한 비트 반전을 수행하여 적대적 견고성의 향상을 달성하고자 한다.

*2.3 혼동 대칭 행렬*

혼동 행렬을 사용하기 위해 대칭 행렬로 변환해 주는 과정이 필요하다. 선행 연구[1]에서는 혼동 대칭 행렬 $\Gamma$를 제안한다. 식(1)의 정의에 따라 $\Gamma$는 대칭 행렬의 성질을 가지며 대각 원소는 0으로 설정한다. $\Gamma$은 각 클래스의 오류 확률을 나타내는 행렬이다. $\Gamma_{ij}$의 값이 크다는 것은 서로의 클래스로 오 분류할 확률이 높다는 것을 의미한다.

$$\Gamma_{ij}=\begin{cases}0, & i=j\\ \dfrac{e_{ij}+e_{ji}}{p_i+p_j}, & i\neq j\end{cases} \tag{1}$$

$$\begin{aligned} e_{ij} &= card(m \mid Y_{true,m}=i,\ Y_{pred,m}=j)\\ p_i &= card(n \mid Y_{true,n}=i). \end{aligned} \tag{2}$$

*2.4 부호어 비트 반전 알고리즘*

클래스 개수를 $M$이라 할 때, 상위 $M$개 부호어를 순서대로 할당한 후 학습을 진행한다. 이후 추론 결과 얻은 혼동 행렬을 통해 클래스 간 구분 가능성을 반영하였다. 만일 어떤 두 클래스 간의 오류 확률이 높다면 비트 반전을 통해 부호어 간의 해밍 거리를 증가시킨다. 비트 반전의 목적은 해밍 거리를 증가시켜 오류 정정 능력을 향상시키는 것이다.

부호어의 길이를 $N=2^k$라 할 때, 가능한 비트 반전의 경우의 수는 $M\times N$이다. 각 반복에서는 단 한 번의 비트 반전을 수행한다. 비트 반전 시 고려할 조건은 다음과 같다.

1. 어떤 두 부호어 간의 해밍 거리도 $2^{k-1}-1$보다 작을 수 없다.
2. 각 반복에서 식(3)을 최대화하는 비트 반전을 수행한다.

식 (3)은 오류 확률이 높은 클래스들에 큰 해밍 거리를 갖는 부호어들이 할당될 경우 최대화된다. 본 연구에서는 적절한 부호어 할당을 찾는 방식으로 최대화하는 대신, 비트 반전을 통해 부호어 간 해밍 거리를 증가시킴으로써 같은 효과를 얻고자 하였다.

해당 알고리즘에서는 비트 반전을 여러 번 수행하는데, 식(3)의 증가량이 수렴할 경우 비트 반전 알고리즘을 종료한다. 비트 반전의 목적은 부호어 간의 해밍 거리를 최적화하여 적대적 견고성을 향상시키는 데 있다.

$$\sum(D\odot\Gamma), \tag{3}$$

*2.6 실험결과*

실험에 사용한 데이터셋은 MNIST, CIFAR10이다. 모델은 TanhEnsemble[4], 코드북은 하다마드 행렬[3]을 사용하였고 PGD 공격의 정확도를 측정하였다. MNIST에 대하여 부호어 길이, 에포크, 배치 크기, 엡실론, pgd 공격 반복 횟수는 각각 16, 150, 100, 0.3, 500으로 설정하였다. CIFAR10에 대하여  부호어 길이, 에포크, 배치 크기, 엡실론, pgd 공격 반복 횟수는 각각 32, 200, 100, 0.031, 200으로 설정하였다.

학습 데이터 전처리 시 적대적 변형을 추가하는 대신 가우시안 잡음과 데이터 증강 기법을 추가하여 적대적 공격에 대한 견고성을 확보하고자 하였다. 비트 반전을 진행하여 코드북을 최적화하고 추론을 진행하였을 때, 적대적 공격에 대한 정확도 향상을 관찰할 수 있었다. MNIST와 CIFAR10 데이터셋 모두 비트 반전을 통해 약 5%의 정확도 향상을 도출하였다. 이는 코드북 최적화가 적대적 견고성 측면에서 중요한 요소라는 것을 시사한다.

| MNIST ($\varepsilon$ = 0.3) | | |
|---|---|---|
| Models | Clean | PGD |
| ECOC | 99.45 | 89.80 |
| **Ours** | 99.20 | **95.65** |

| CIFAR10 ($\varepsilon$ = 0.031) | | |
|---|---|---|
| Models | Clean | PGD |
| ECOC | 77.70 | 58.25 |
| Ours | 74.75 | **62.64** |

표 1 MNIST, CIFAR10 ECOC 모델 정확도

## Ⅲ. 결 론

클래스 간의 구분 가능성을 고려한 부호어 비트 반전 알고리즘은 채널 잡음과 유사한 적대적 공격을 방어하며 적대적 견고성을 향상시켰다. 본 논문에서는 혼동 대칭 행렬과 거리 행렬의 하다마드 곱(3)을 최대화하는 방향으로 비트 반전을 수행하였다. 실험 결과 비트 반전 전과 후에서 유의미한 정확도의 차이를 관찰할 수 있었고, 최적에 가까운 코드북을 찾는 데 성공하였다. 클래스 간 구분 가능성을 고려하여 비트 반전을 하였기 때문에 모델 구조를 단순화할 수 있었으며, 정확도의 향상 또한 입증할 수 있었다.

추후 BCH code, 저밀도 패리티 검사 코드 등의 다른 오류 정정 부호를 적용하거나, 모델의 구조를 변형해 적대적 견고성을 향상하는 등의 연구가 가능할 것으로 보인다.


## ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021001016004, 6G/B5G xURLLC를 위한 유연한 신뢰도의 채널코딩), (No.RS-2021-II212068, 인공지능 혁신 허브 연구 개발)과 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업(차세대통신)의 연구 결과입니다.


## 참 고 문 헌


[1] Li Wan, Tansu Alpcan, Emanuele Viterbo, and Margreta Kuijper. Efficient error-correcting output codes for adversarial learning robustness. In ICC 2022-IEEE Inter- national Conference on Communications, pages 2345 - 2350. IEEE, 2022.

[2] DIETTERICH, Thomas G.; BAKIRI, Ghulum. Solving multiclass learning problems via error-correcting output codes. Journal of artificial intelligence research, 1994, 2: 263-286.

[3] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of machine learning research, 1(Dec):113 - 141, 2000.

[4] VERMA, Gunjan; SWAMI, Ananthram. Error correcting output codes improve probability estimation and adversarial robustness of deep neural networks. Advances in Neural Information Processing Systems, 2019, 32.