개별연구 (CSC AI모델 보안 강화 방법 연구)

이름 : 김동현
학과 : 컴퓨터AI학부
학번 : 2021111982
담당 교수 : 손윤식 교수님

1. 연구 배경 및 목표

인공지능(AI) 기술은 의료 진단, 자율주행 시스템, 금융 분석 등 다양한 분야에서 핵심적인 의사결정 도구로 활용되며, 현대 사회의 필수 기술로 자리 잡고 있다. 특히 의료 및 교통과 같이 인간의 생명과 안전에 직접적인 영향을 미치는 영역에서는 AI 시스템의 판단이 높은 정확도(Accuracy)뿐만 아니라, 예측 불가능한 환경 변화나 외부 위협에도 안정적으로 대응할 수 있는 신뢰성(Reliability)과 안전성(Safety)을 요구받고 있다.

이에 따라 AI 모델의 성능 평가는 단순한 분류 정확도에 국한되지 않고, 실제 환경에서 발생할 수 있는 다양한 교란 상황에서도 일관된 성능을 유지할 수 있는지에 대한 종합적인 검증의 필요성이 점차 강조되고 있다.

이러한 요구에도 불구하고, 최근 AI 시스템의 신뢰성을 근본적으로 위협하는 문제로 적대적 공격(Adversarial Attack)이 주목받고 있다. 적대적 공격이란 인간이 인지하기 어려운 수준의 미세한 노이즈나 의도적인 교란을 입력 데이터에 추가함으로써, AI 모델의 예측 결과를 왜곡시키는 공격 기법을 의미한다.

적대적 공격의 위험성은 실제 적용 사례를 통해 확인되고 있다. 자율주행 환경에서는 도로 표지판에 특정 패턴의 스티커가 부착될 경우, AI가 '정지' 표지판을 다른 표지로 오인식하여 심각한 교통사고로 이어질 수 있다. 또한 의료 영상 분석 분야에서는 극히 미세한 노이즈가 추가된 영상으로 인해 악성 종양을 정상 조직으로 오판독하는 치명적인 오류가 발생할 가능성도 보고되고 있다. 이와 같은 사례들은 AI 시스템이 높은 정확도를 보이더라도, 외부 공격이나 환경 변화에 취약할 경우 실제 서비스 환경에서 심각한 위험 요소로 작용할 수 있음을 시사한다. 따라서 AI 모델의 성능 평가는 정확도뿐만 아니라, 적대적 공격에 대한 강건성(Robustness)을 함께 고려하는 방향으로 확장될 필요가 있다.

본 연구는 다양한 적대적 공격 환경에서 AI 모델의 강건성(Robustness)과 일반 성능(Clean Accuracy) 간의 상충 관계를 체계적으로 분석하는 것을 목표로 한다. 본 연구에서 강건성이란, 입력 데이터에 제한된 크기의 교란이 가해졌을 때에도 모델이 예측 성능을 안정적으로 유지하는 능력을 의미한다. 이를 정량적으로 평가하기 위해, 아래 실험에서는 적대적 공격 하에서의 분류 정확도(Robust Accuracy)를 주요 지표로 사용한다. 이러한 분석을 통해 단순히 강건성만을 극대화하는 접근이 아니라, 원본 데이터에 대한 분류 정확도와 적대적 환경에서의 강건성 간의 균형을 고려한 보안 모델 설계 방향을 제시하고자 한다.

2. 실험 설정 및 이론적 배경

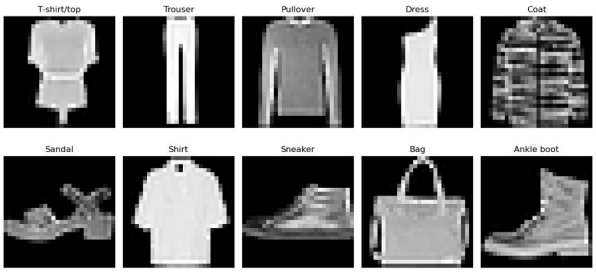
2-1. 실험 환경

- Operating System: Windows 11 (WSL2 기반 Linux 환경)
- Python: 3.13.9 (Miniconda 환경)
- Deep Learning Framework: TensorFlow (GPU 지원)
- Adversarial Attack Library: IBM Adversarial Robustness Toolbox (ART)
- GPU: NVIDIA GeForce RTX 4060 Ti
- CUDA: NVIDIA CUDA Toolkit (WSL2 지원 환경)

2-2. 사용 데이터셋

: Fashion MNIST

Fashion MNIST는 기존의 손글씨 숫자 데이터셋인 'Classic MNIST'를 대체하기 위해 Zalando Research에서 공개한 이미지 데이터셋이다. 0부터 9까지의 숫자가 아닌, 10가지 종류의 패션 아이템(의류, 신발, 가방 등)으로 구성되어 있다.



- 이미지 크기: 28 x 28 픽셀 (Pixel)
- 채널: 흑백 (Grayscale, 1 Channel)
- 데이터 개수: 총 70,000장
- 학습용(Training set): 60,000장 (default 값)
- 테스트용(Test set): 10,000장 (default 값)
- 클래스(Label): 총 10개 (0~9)

라벨 (Label)	클래스 (Class)	설명
0	T-shirt/top	티셔츠/상의
1	Trouser	바지
2	Pullover	스웨터(풀오버)
3	Dress	원피스
4	Coat	코트
5	Sandal	샌들
6	Shirt	셔츠 (단추가 있는 형태)
7	Sneaker	운동화
8	Bag	가방
9	Ankle boot	발목 부츠

- 본 연구에서의 선정 이유

본 연구에서는 실험에 가장 적합한 데이터셋을 선정하기 위해 대표적인 공개 데이터셋인 MNIST와 CIFAR-10을 후보로 선정하여 사전 분석을 진행하였다.

우선, MNIST 데이터셋은 흑백 숫자 이미지로 구성되어 있어 구조가 지나치게 단순하다는 한계가 있었다. 실험 결과, MNIST는 아주 미세한 노이즈(공격)만 추가해도 분류 결과가 급격하게 바뀌는 등 민감도가 높게 나타났다. 이는 데이터 자체가 너무 단순하여 발생하는 현상으로, 공격과 방어 기법의 실제 성능을 정확하게 측정하기에는 결과가 왜곡될 가능성이 있다고 판단하였다.

반면, CIFAR-10 데이터셋은 컬러 이미지인 만큼 현실적이고 복잡도가 높다는 장점이 있다. 하지만 공격 데이터를 생성하거나 방어 기법(Adversarial Training)을 적용하는 과정에서 연산량이 급증하는 문제가 있었다. 한정된 기간과 컴퓨팅 환경에서 다양한 파라미터를 바꿔가며 반복 실험을 진행하기에는 시간적 효율성이 낮다고 판단하였다.

따라서 본 연구에서는 계산 효율성과 데이터 난이도의 균형을 맞춘 Fashion-MNIST를 최종적으로 선정하였다. Fashion-MNIST는 MNIST와 같은 크기(28x28)라 연산 속도가 빠르면서도, 의류라는 이미지 특성상 형태가 다양해 적절한 난이도를 가지고 있다. 이를 통해 효율적이면서도 MNIST보다 훨씬 신뢰성 있는 연구를 수행하고자 하였다.

2-3. 공격 기법

모델의 취약점을 분석하고 방어 성능을 검증하기 위해 적대적 공격(Adversarial Attack) 기법 중 대표적인 회피 공격(Evasion Attack) 기법인 FGSM을 실험에 적용하였다. 이는 모델의 손실 함수의 기울기 정보를 악용하는 공격 방식이다.

- **FGSM (Fast Gradient Sign Method)**

FGSM은 Goodfellow 등(2014)이 제안한 기법으로, 입력 데이터에 대해 모델의 손실이 증가하는 방향(기울기의 부호, Sign)으로 미세한 노이즈를 단 한 번(One-shot) 추가하여 적대적 예제를 생성하는 방법이다. 연산 비용이 낮아 적대적 학습을 위한 데이터를 대량으로 생성할 때 효율적이지만, 공격의 정교함은 다소 떨어진다는 특징이 있다. 수식적으로는 원본 이미지에 입실론(ε) 크기만큼의 기울기 부호 값을 더해 이미지를 왜곡시킨다.

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

2-4. 방어 기법

모델의 강건성(Robustness)을 확보하기 위해 적대적 학습(Adversarial Training) 기법을 적용하였다. 적대적 학습이란 학습 과정에서 원본 데이터뿐만 아니라, 모델을 의도적으로 속이기 위해 생성된 적대적 예제(Adversarial Examples)를 훈련 데이터에 함께 포함시켜 학습하는 방법이다. 이는 마치 바이러스 백신을 맞듯, 모델이 노이즈가 포함된 데이터의 분포를 미리 경험하고 학습하게 함으로써 실제 공격 상황에서의 오분류를 방지하는 것을 목표로 한다.

노이즈가 포함된 데이터(적대적 예제)를 생성하기 위해 위에서 설명한 FGSM 방식을 사용할 예정이다.

2-5. 사용 라이브러리

구체적인 구현을 위해 Python 기반의 **ART(Adversarial Robustness Toolbox)** 라이브러리를 활용하였다. 주요 구성 요소로는 공격 데이터를 생성하는 FastGradientMethod와 이를 이용해 방어 모델을 훈련시키는 AdversarialTrainer를 사용하였으며, 각 클래스의 역할과 주요 파라미터는 다음과 같다.

1) FastGradientMethod (공격 데이터 생성)

적대적 학습에 필요한 공격 데이터를 생성하는 핵심 모듈이다. 이 함수는 방어 대상 모델이 정답을 맞추기 가장 어렵게 만드는 방향(기울기)을 계산한 뒤, 원본 이미지에 미세한 노이즈를 추가하는 역할을 한다. 주요 파라미터는 다음과 같다.

- **estimator** : 공격 대상이 되는 모델 객체를 의미한다.
- **eps (Epsilon)**: 이미지에 추가할 노이즈의 최대 크기(강도)를 의미한다.

2) AdversarialTrainer (방어 학습 수행)

앞서 정의한 공격 모듈을 이용해 실제 방어 학습을 진행하는 클래스이다. 학습이 진행되는 동안 실시간으로 공격 데이터를 생성하고, 이를 모델 가중치 업데이트에 즉시 반영하여 모델의 방어력을 높인다. 주요 파라미터는 다음과 같다.

- **classifier** : 학습을 수행할 모델 객체를 의미
- **attacks** : 학습에 사용할 공격 기법을 지정한다. 위에서 정의한 FastGradientMethod 객체가 입력되며, 이를 통해 적대적 예제를 실시간으로 생성한다.
- **ratio** : 학습 데이터 한 배치(Batch) 내에서 '적대적 예제'가 차지하는 비율을 의미한다. 만약 0.5로 설정한 경우, 원본 데이터와 적대적 예제 데이터가 50:50으로 섞여서 학습된다.

3. 실험 과정

3-1. 초기 모델(Baseline) 설계 및 학습

- 초기 모델 구조

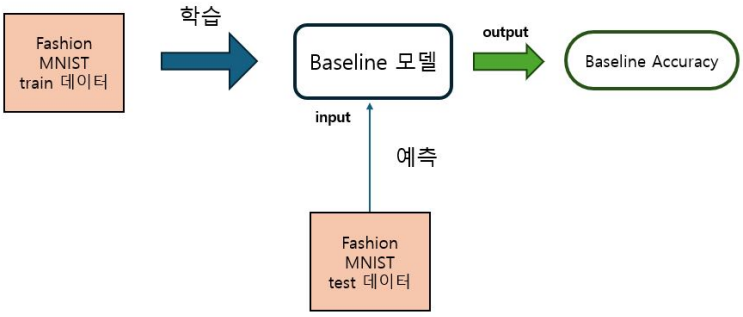
```
def create_cnn_model():
    model = Sequential([
        # 입력층 (28x28x1)
        Input(shape=(28, 28, 1)),

        # 블록 1
        Conv2D(32, kernel_size=(3, 3), padding='same'),
        Activation('relu'),
        MaxPooling2D(pool_size=(2, 2)),

        # 블록 2
        Conv2D(64, kernel_size=(3, 3), padding='same'),
        Activation('relu'),
        MaxPooling2D(pool_size=(2, 2)),

        # 분류기 (Classifier) 부분
        Flatten(),

        #중간 은닉층
        Dense(128, activation='relu'),

        # 과적합 방지 드롭아웃
        Dropout(0.3),

        # 최종 출력층 (10개 클래스)
        Dense(10, activation='softmax')
    ])

    # Adam 옵티마이저 (학습률 0.001)
    model.compile(loss=categorical_crossentropy,
                  optimizer=Adam(learning_rate=0.0001),
                  metrics=['accuracy'])

    return model
```

Fashion MNIST 데이터셋의 특성을 고려하여, 2단계의 합성곱 블록(Convolutional Block)과 1단계의 은닉층(Hidden Layer)을 갖춘 기본적인 CNN 모델을 설계하였다.

첫 번째 블록(32 filters)과 두 번째 블록(64 filters)은 이미지의 저수준 특징(Edge)에서 고수준 특징(Shape)으로 이어지는 계층적 특징을 추출하며, MaxPooling을 통해 주요 특징을 압축하고 연산 효율을 높였다. 특히 padding='same'을 적용하여 이미지 경계의 정보 손실을 최소화하였다. 추출된 특징 맵은 Flatten을 거쳐 128개의 노드를 가진 완전 연결층(Dense Layer)으로 전달되어, 시각적 패턴 간의 비선형적 관계를 학습한다. 마지막으로 과적합 방지를 위해 Dropout(0.3)을 적용한 후, Softmax 활성화 함수를 통해 10개 클래스에 대한 최종 분류 확률을 출력하도록 구성하였다.

- 초기 모델 학습

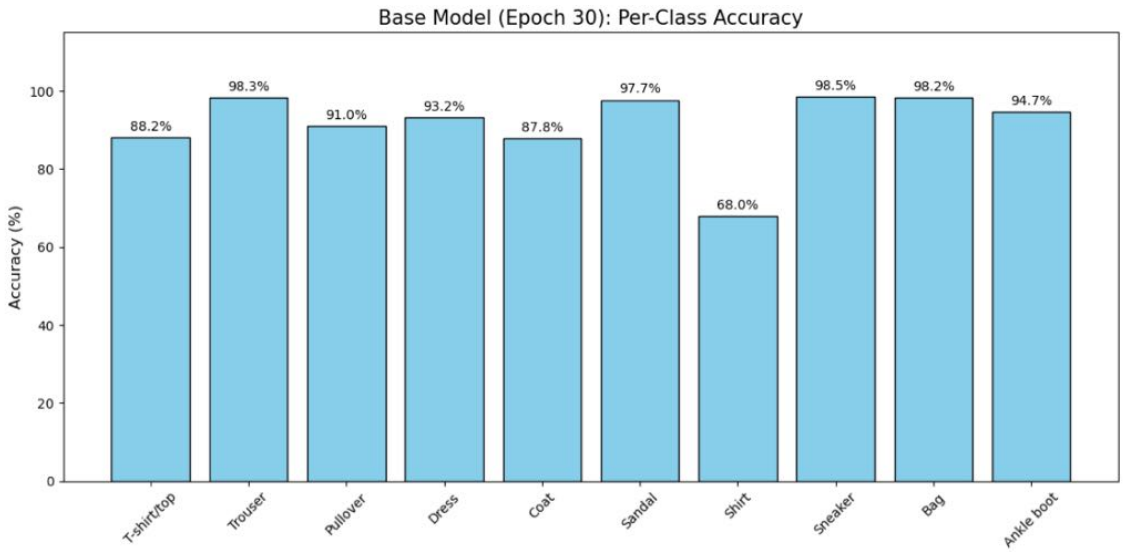
먼저 앞서 설계한 CNN 모델을 사용하여 Fashion-MNIST 데이터셋에 대한 기본 분류 성능을 먼저 측정하였다. 이 단계의 목적은 이후 공격 및 방어 실험의 기준이 될 초기 모델(Baseline model)을 설정하는 것이다.



모델 성능 평가지표로는 Accuracy를 사용하였다. Fashion-MNIST 데이터셋은 각 클래스가 동일한 수의 샘플로 구성된 균형 잡힌 데이터셋이기 때문에, Accuracy는 F1-score 등 다른 분류 성능 지표와 유사한 경향을 보였다. 또한 Accuracy는 가장 직관적이고 보편적으로 사용되는 지표이므로, 성능 비교를 위한 기준 지표로 적합하다고 판단하였다. 향후 보다 정밀한 성능 분석이 필요할 경우, 추가적인 성능 지표를 활용한 분석을 할 예정이다.

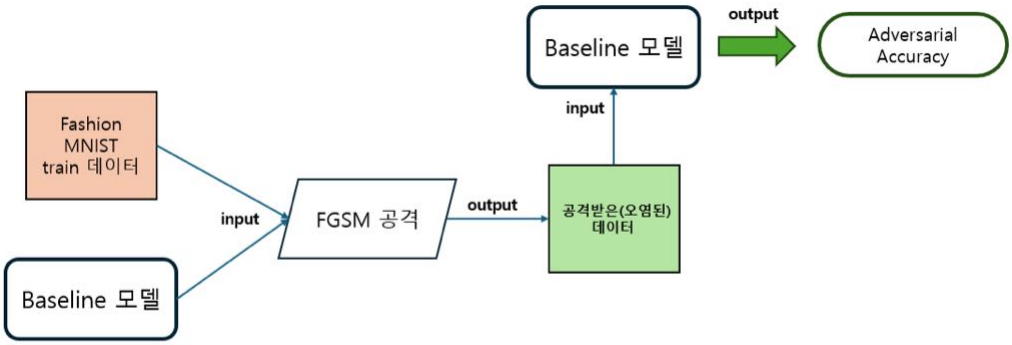
학습 에포크(epoch)의 영향을 분석하기 위해, 에포크 수를 5부터 30까지 5 단위로 증가시키며 모델을 학습하고 성능을 측정하였다. 실험 결과는 다음과 같다.

```
>>> 에포크별 성능 측정 (5 ~ 30)...
Testing Epoch 5... -> Accuracy: 87.61%
Testing Epoch 10... -> Accuracy: 89.25%
Testing Epoch 15... -> Accuracy: 89.69%
Testing Epoch 20... -> Accuracy: 90.79%
Testing Epoch 25... -> Accuracy: 91.03%
Testing Epoch 30... -> Accuracy: 91.56%

==================================================
 Best Result: Epoch 30 (Accuracy: 91.56%)
==================================================
```



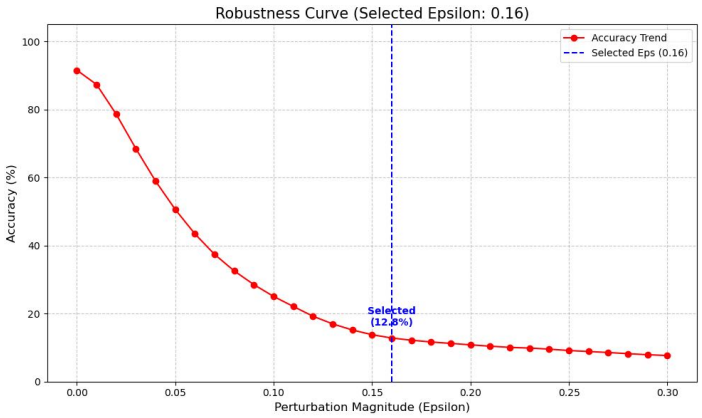
에포크가 증가함에 따라 전반적으로 Accuracy가 향상되는 경향을 보였으며, 에포크 30에서 가장 높은 Accuracy(91.56%)를 기록하였다. 이에 따라, 에포크 30에서 학습된 모델을 본 연구의 Baseline 모델로 설정하였다. 또한, 각 클래스 별 Accuracy를 비교했을 때, shirt 클래스가 68%로 가장 낮고, sneaker 클래스가 98.5%로 가장 높은 것을 볼 수 있었다.

3-2. 초기 모델에 대하여 공격 수행



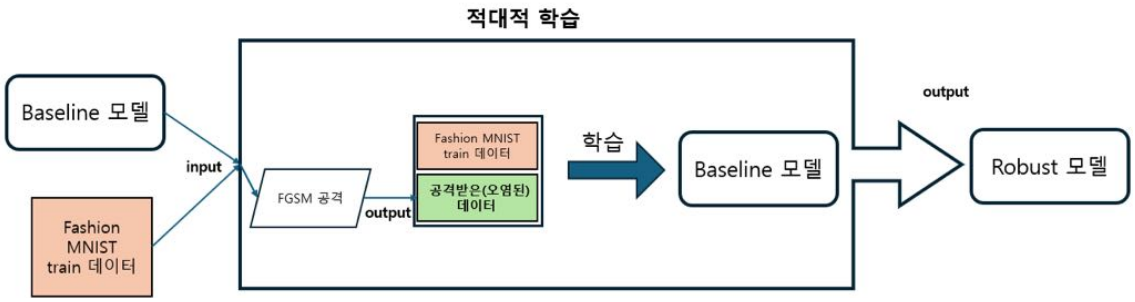
설정된 Baseline 모델의 취약성을 확인하기 위해, FGSM 공격 기법을 사용하여 적대적 공격을 수행하였다. 공격 강도를 조절하는 파라미터인 ε(epsilon) 값을 0.01부터 0.30까지 0.01 단위로 증가시키며 공격을 적용하였고, 각 ε 값에 대해 모델의 분류 정확도를 측정하였다.

```
>>> Epsilon 값에 따른 accuracy 변화 측정
기준 정확도: 91.56%
-----------------------------------------------------------------
Epsilon    | Accuracy    | Drop(Slope)      | Status
-----------------------------------------------------------------
0.01       | 87.36    % | -4.20         % |
0.02       | 78.62    % | -8.74         % |
0.03       | 68.47    % | -10.15        % |
0.04       | 58.94    % | -9.53         % |
0.05       | 50.60    % | -8.34         % |
0.06       | 43.42    % | -7.18         % |
0.07       | 37.38    % | -6.04         % |
0.08       | 32.52    % | -4.86         % |
0.09       | 28.49    % | -4.03         % |
0.10       | 25.03    % | -3.46         % |
0.11       | 22.09    % | -2.94         % |
0.12       | 19.22    % | -2.87         % |
0.13       | 16.96    % | -2.26         % |
0.14       | 15.17    % | -1.79         % |
0.15       | 13.79    % | -1.38         % |
0.16       | 12.81    % | -0.98         % | Optimal Epsilon
0.17       | 12.17    % | -0.64         % |
0.18       | 11.65    % | -0.52         % |
0.19       | 11.23    % | -0.42         % |
0.20       | 10.81    % | -0.42         % |
0.21       | 10.41    % | -0.40         % |
0.22       | 10.07    % | -0.34         % |
0.23       | 9.86     % | -0.21         % |
0.24       | 9.54     % | -0.32         % |
0.25       | 9.14     % | -0.40         % |
0.26       | 8.85     % | -0.29         % |
0.27       | 8.56     % | -0.29         % |
0.28       | 8.21     % | -0.35         % |
0.29       | 7.91     % | -0.30         % |
0.30       | 7.66     % | -0.25         % |
-----------------------------------------------------------------
-> 최종 선택 Epsilon: 0.16
-> FGSM 공격에 대한 baseline 모델의 accuracy : 12.809999999999999%
```



실험 결과, ε 값이 증가함에 따라 모델의 분류 정확도는 전반적으로 감소하는 경향을 보였다. 초기 구간에서는 ε 증가에 따라 Accuracy가 급격히 감소하였으나, 일정 수준 이후에는 연속된 ε 값 사이의 Accuracy 감소폭이 점차 완만해졌다. 본 연구에서는 연속된 ε 값 간 Accuracy 감소율이 1% 미만으로 낮아지는 지점을 공격 효과가 포화(saturation)에 도달한 시점으로 판단하였고, 이에 따라 ε = 0.16을 FGSM 공격에 대한 최적의 ε 값으로 선정하였고 이때의 Accuracy는 12.8%이다. 이후 모든 공격 및 방어 실험에서는 동일한 ε 값을 사용하여 실험 조건의 일관성을 유지하였다.

3-3. 적대적 학습 수행 (방어 기법 적용) 및 원본 데이터에 대한 성능 평가



(1) 적대적 학습 대상 모델 비교

먼저, 적대적 학습을 수행하는 대상 모델에 따른 성능 차이를 분석하였다. 이를 위해 "기존 Baseline 모델을 기반으로 적대적 학습을 수행한 경우", "동일한 구조의 새로운 모델을 초기화한 후 적대적 학습을 수행한 경우" 두 가지 경우를 비교하였다. 파라미터는 Epsilon=0.16, Ratio=0.5으로 설정하고 실험을 진행하였다.

```
======================================================================
Strategy                                  | Robust Accuracy
----------------------------------------------------------------------
기존 학습된 모델에 추가 적대적 학습 (Pre-trained)       | 84.34%
새로운 모델에 적대적 학습 (New Model)                | 77.42%
======================================================================
```

실험 결과, Baseline 모델을 기반으로 적대적 학습을 수행한 경우(Pre-trained)가 새로운 모델에서 시작한 경우보다 약 7% 높은 Accuracy를 보였다. 이는 사전에 학습된 특징 표현을 기반으로 적대적 학습을 수행하는 것이 모델의 강건성 향상에 보다 효과적임을 의미한다. 따라서 이후의 모든 적대적 학습 실험은 Baseline 모델을 기반으로 진행하였다.

(2) 공격 데이터에 따른 성능 비교

다음으로, 적대적 학습을 수행한 모델에 대해 새롭게 생성한 FGSM 공격 데이터와 기존 3-2 단계에서 사용한 공격 데이터에 대한 성능을 비교하였다. 마찬가지로 파라미터는 Epsilon=0.16, Ratio=0.5, epochs=10 으로 설정하고 실험을 진행하였다.

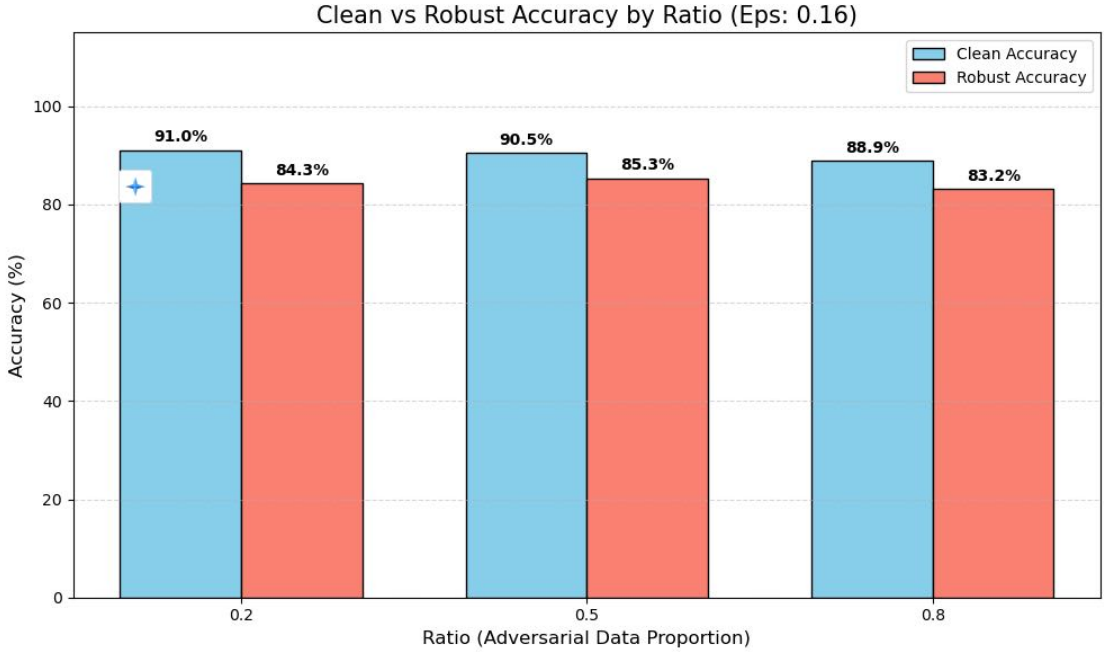
```
학습 설정: Epsilon=0.16, Ratio=0.5, Epochs=10
공격 데이터에 따른 성능 비교
Precompute adv samples: 100% 1/1 [00:00<00:00, 192.28it/s]
Adversarial training epochs: 100% 10/10 [09:42<00:00, 58.17s/it]
>>> 1. 방어된 모델에 대한 '새로운 공격(Re-attack)' 시도
   -> 새로운 공격에 대한 모델의 성능 = 85.63%

>>> 2. 저장된 '이전 공격 데이터(final_adv_data)'에 대한 모델의 성능 측정
   -> 저장된 적대적 예제(과거의 공격)에 대한 정확도 = 89.20%
```

실험 결과, 새로운 FGSM 공격에 대해 측정한 Accuracy가 기존 공격 데이터에 대한 Accuracy보다 약 4% 더 낮게 나타났다. 이는 기존 공격 데이터가 적대적 학습 과정 중 일부 포함되어 학습되었기 때문에, 모델이 해당 공격 패턴에 대해 상대적으로 더 잘 적응했기 때문으로 해석할 수 있다. 반면, 새롭게 생성된 공격 데이터는 학습에 포함되지 않은 공격으로, 모

델의 일반적인 강건성을 보다 엄격하게 평가하는 지표로 작용한다고 볼 수 있다.

(3) 적대적 학습 비율(ratio)에 따른 비교 실험

적대적 학습 시 배치 내에서 적대적 예제가 차지하는 비율(ratio)이 모델 성능에 미치는 영향을 분석하기 위해,ratio 값을 0.2, 0.5, 0.8로 설정하여 비교 실험을 수행하였다. 모든 실험에서 공격 기법은 FGSM으로 통일하였으며, 각 ratio 값에 대해 원본(clean) 데이터에 대한 분류 정확도, 적대적 공격 데이터에 대한 분류 정확도 두 가지 성능을 함께 측정하였다. 이를 통해 적대적 학습 비율 변화에 따른 강건성 향상과 원본 성능 저하 간의 trade-off를 분석하였다.

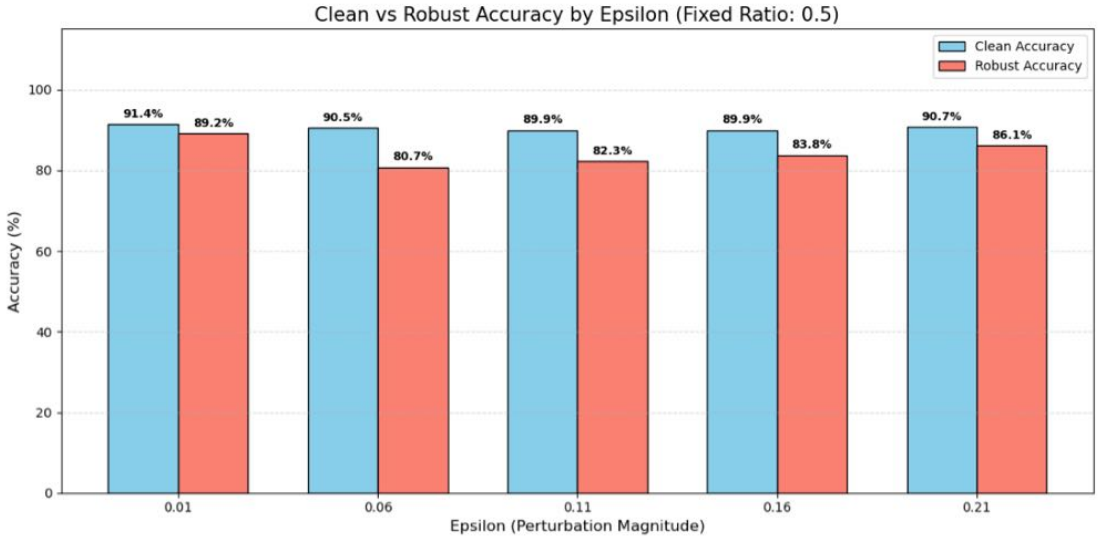


Epsilon = 0.16 조건에서 수행한 비교 실험 결과, 적대적 데이터의 비율이 50%(Ratio 0.5)일 때 Clean Accuracy(90.5%)와 Robust Accuracy(85.3%) 간의 균형이 가장 이상적으로 나타났다. Ratio가 0.8로 과도하게 높아질 경우, 원본 데이터에 대한 정확도 하락폭이 커질 뿐만 아니라 방어 성능 또한 0.2 수준보다 낮은 83.2%로 떨어지는 현상이 관찰되었다. 이는 학습 데이터 내 노이즈 비율이 임계점을 넘으면 모델의 수렴을 방해하여 전반적인 성능 저하를 유발하기 때문으로 판단된다. 결론적으로, 적절한 방어 성능 확보와 일반화 능력 유지를 위해서는 Ratio=0.5가 최적의 하이퍼파라미터로 볼 수 있다.

(4) 적대적 학습 시 공격 강도(입실론)에 따른 비교 실험

앞선 실험을 통해 적대적 데이터 혼합 비율(Ratio)은 0.5가 Clean Accuracy(정상 데이터 정확도)와 Robust Accuracy(방어 정확도) 간의 트레이드오프(Trade-off)를 최소화하는 최적점임을 확인하였다.

이에 본 실험에서는 Ratio를 0.5로 고정한 상태에서, 공격의 강도를 결정하는 Epsilon 값을 0.01에서 0.21까지 단계적으로 증가시키며 모델의 방어 성능 변화를 분석하였다. 평가 시에는 학습에 사용된 Epsilon과 동일한 강도의 공격을 가하여 모델의 적응력을 검증하였다.



실험 결과, 원본 데이터에 대한 분류 정확도(Clean Accuracy)는 적대적 학습 과정에서 사용된 적대적 예제를 생성하기 위한 공격 강도 ε가 0.01에서 0.21까지 증가하는 환경에서도 약 89.9%~91.4% 범위를 유지하며 전반적으로 매우 안정적인 성능을 보였다. 이는 ratio = 0.5로 설정한 적대적 학습이 강한 교란 환경에서도 모델이 원본 데이터의 핵심 특징을 효과적으로 보존하도록 학습되었음을 의미한다. 특히, 원본 데이터로 충분히 학습된 Baseline 모델을 기반으로 적대적 학습을 수행했기 때문에 Clean Accuracy가 안정적으로 유지된 것으로 해석할 수 있다.

한편, 적대적 공격 데이터에 대한 분류 정확도(Robust Accuracy)는 적대적 학습에 사용된 ε 값에 따라 보다 뚜렷한 변화를 보였다. 상대적으로 작은 노이즈를 학습한 경우(ε = 0.01)에는 89.2%의 높은 방어 성능을 보였으나, 보다 큰 교란이 학습에 본격적으로 포함되는 구간(ε = 0.06)에서 80.7%로 일시적인 성능 저하가 관찰되었다. 이후 ε 값이 증가함에 따라 모델이 강한 교란 패턴에 점진적으로 적응하면서 Robust Accuracy는 다시 상승하였고, 가장 큰 노이즈를 학습한 조건인 ε = 0.21에서는 86.1%까지 회복되었다.

종합적으로, 본 실험 결과는 적대적 학습이 원본 데이터에 대한 성능 저하를 최소화하는 동시에, 강한 적대적 공격 환경에서도 의미 있는 수준의 방어 성능을 확보할 수 있음을 보여준다. 이는 제안한 설정이 Clean Accuracy와 Robust Accuracy 간의 균형(trade-off)을 효과적으로 달성했음을 시사한다.

4. 결론

본 연구에서는 적대적 공격 환경에서 AI 분류 모델의 성능 저하 양상을 분석하고, 적대적 학습을 통한 방어 기법이 모델의 강건성(Robustness)과 일반 성능(Clean Accuracy)에 미치는 영향을 실험적으로 평가하였다. 먼저 공격 기법에 대한 실험 결과, FGSM 공격 강도(ε)가 증가함에 따라 분류 정확도가 급격히 감소하는 경향을 확인하였으며, Accuracy 감소율이 포화(Saturation)되는 지점을 기준으로 ε = 0.16을 본 연구의 최적 공격 강도로 선정하였다. 이후 모든 적대적 학습 실험은 해당 ε 값을 기준으로 수행하여 실험 조건의 일관성을 유지하였다.

적대적 학습의 효율성을 분석한 결과, 기존에 학습된 Baseline 모델을 기반으로 방어 학습을 수행한 경우(Pre-trained)가 초기화된 모델을 처음부터 학습한 경우보다 약 7% 높은 정확도를 보였으며, 이는 사전 학습된 특징 표현이 적대적 환경에서도 효과적인 방어 성능 확보에 기여함을 의미한다. 또한 적대적 예제 혼합 비율(Ratio)에 따른 비교 실험에서는 Ratio = 0.5 설정에서 Clean Accuracy 90.5%, Robust Accuracy 85.3%를 기록하여, 일반 성능과 강건성 간의 균형이 가장 우수한 것으로 나타났다.

마지막으로 적대적 학습 시 적용하는 공격 강도(ε)에 따른 성능 변화를 분석한 결과, ε = 0.01 조건에서 원본 데이터에 대한 정확도 91.4%와 적대적 공격 데이터에 대한 정확도 89.2%를 기록하며 가장 높은 성능을 보였다. 이는 초기 Baseline 모델의 성능(91.56%)과 거의 유사한 수준으로, 원본 성능의 손실을 최소화하면서도 의미 있는 방어 성능을 확보한 결과로 해석할 수 있다. 종합적으로, 본 연구에서 제안한 적대적 학습 설정은 성능 저하를 최소화하는 동시에 강건성을 효과적으로 향상시키는 보안 모델이라 평가할 수 있다.

```
============================================================
           Fashion MNIST FGSM 실험 결과
============================================================
           Step          Data Type      Model Used  Accuracy (%)
0   1. Baseline        Clean Image  Standard Model         91.56
1     2. Attack  Adversarial Image  Standard Model         12.81
2    3. Defence  Adversarial Image  Defended Model         89.20
3  4. Trade-off        Clean Image  Defended Model         91.40
```

다만 본 연구는 제한된 컴퓨팅 환경, 단일 데이터셋(Fashion-MNIST), 제한된 모델 구조(CNN), 그리고 FGSM 공격 기법을 중심으로 수행되었다는 한계를 가진다. 따라서 제안된 최적 하이퍼파라미터 설정은 본 연구의 실험 환경에 한정된 결과이며, 향후 연구에서는 보다 다양한 데이터셋과 공격 기법(PGD 등), 모델 구조를 확장 적용함으로써 결과의 일반성을 검증할 필요가 있다.

부록. 실험 코드

본 연구에서 사용된 모든 실험 코드는 공개 GitHub 저장소를 통해 제공한다.
코드에는 데이터 전처리, CNN 모델 학습, FGSM 공격 생성, 적대적 학습 및 성능 평가 과정이 포함되어 있으며, 실험 결과의 재현성을 확보하였다.
https://github.com/kddhhh23/adversarial-robustness-study.git

참고 자료

박재경, 장준서. (2023-01-12). AI 모델의 적대적 공격 대응 방안에 대한 연구. 한국컴퓨터정보학회 학술발표논문집 , 대전.
김규형. "이미지 데이터 변조의 적대적 공격에 대한 CNN과 SNN의 강건성 비교연구." 국내석사학위논문 아주대학교 일반대학원, 2024. 경기도