정 보 보 호 학 회 지
제 30 권 제6호, 2020. 12

이미지 기반 적대적 사례 생성 기술 연구 동향

오 희 석*

요 약

다양한 응용분야에서 심층신경망 기반의 학습 모델이 앞 다투어 이용됨에 따라 인공지능의 설명 가능한 동작 원리 해석과, 추론이 갖는 불확실성에 관한 분석 또한 심도 있게 연구되고 있다. 이에 심층신경망 기반 기계학습 모델의 취약성이 수면 위로 드러났으며, 이러한 취약성을 이용하여 악의적으로 모델을 공격함으로써 오동작을 유도하고자 하는 시도가 다방면으로 이루어짐에 의해 학습 모델의 강건함 보장은 보안 분야에서의 쟁점으로 부각되고 있다. 모델 추론의 입력으로 이용되는 이미지에 교란값을 추가함으로써 심층신경망의 오분류를 발생시키는 임의의 변형된 이미지를 적대적 사례라 정의하며, 본 논문에서는 최근 인공지능 및 컴퓨터비전 분야에서 이루어지고 있는 이미지 기반 적대적 사례의 생성 기법에 대하여 논한다.

Ⅰ. 서 론

2012년 이미지 인식 대회 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)에서 CNN(Convolutional Neural Network)기반의 심층신경망 구조 AlexNet[1]이 소개되며 기존 객체 분류 기술 대비 오차를 10% 이상 감소시키는 압도적인 결과를 보여줌에 따라, 특정 도메인에서의 전문 지식을 통해 수작업으로 이루어지던 특징(feature)의 추출은 대규모의 데이터로부터 자동으로 도출되는 심층신경망 기반의 표현력 학습 방법으로 그 패러다임이 변화하였다. 급기야 2015년 ResNet[2]의 소개와 함께 이미지 인식 분야에서는 사람의 객체 분류 능력을 뛰어 넘는 성능 결과가 공개되었고, 이후 컴퓨터비전 분야를 비롯한 이미지처리, 그래픽스 및 자연어처리 등의 다양한 영역에서 심층신경망 기반의 기계학습 기술 활용은 데이터 분석에 있어 선택이 아닌 필수 사항으로 자리매김하였다.

하지만 심층신경망 기반의 추론은 학습된 모델이 어떠한 물리적 의미를 표현하는 특징을 추출하였는지 설명이 어려운 블랙박스라는 단점을 가지며, 또한 판단이나 의사 결정에 필요한 적절한 정보의 부족으로 잘못된 판단을 내릴 수 있는 불확실성(uncertainty)을 내재하고 있다[3]. 이러한 불확실성은 심층신경망 기반으로 학습된 대부분의 최첨단 기술들이 악의적 의도를 가진 공격에 노출되었을 시, 적용 분야에 따라서는 지극히 위험한 판단을 내릴 가능성을 내포하고 있다는 것을 의미하며 실제로 Garg등은 도로 위 차량 정지 표지판에 작은 메모지를 붙이는 것만으로도 95%의 정확도를 자랑하는 인공지능 인식 모듈이 이를 45km/h 이하 속도 주행 가능 표지판으로 인식해 오작동하는 사례를 보여주었다[4].

설명 가능한(explainable) 인공지능을 위한 기술 연구와 예측 모델이 갖는 불확실성에 대한 심도 있는 분석이 수행됨에 따라 심층신경망 기반 기계학습의 취약성이 수면 위로 드러나게 되었고, 이에 대한 산업계와 학계의 관심은 갈수록 증가하고 있는 추세이다. 이에 국방, 의료, 감시 시스템을 비롯한 인공지능 기술을 활용한 각종 첨단 응용분야에서의 보안 이슈가 새롭게 대두되고 있으며 최근 심층신경망 모델에 대한 공격과 방어 관련 다양한 연구가 수행되고 있다[5].

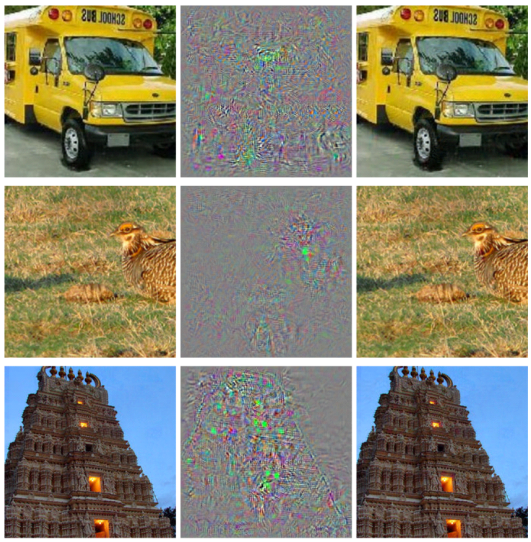
본 논문에서는 최근 컴퓨터비전 및 인공지능 학회에서 발표되는 심층신경망 모델들의 취약성을 이용한 다양한 이미지 기반 공격 방법에 대한 기술적 특징을 살펴보고 연구 동향을 분석한다.

본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1G1A110067411).
* 한성대학교 IT융합공학부 (조교수, ohhs@hansung.ac.kr)

Ⅱ. 적대적 사례

2.1. 적대적 사례의 정의

Szegedy등은 심층신경망 기반 이미지 분류 모델의 흥미로운 맹점을 보고하였다. 육안으로 인지 불가능한 아주 작은 크기의 잡음과 같은 교란값(perturbation)을 이미지에 추가하였을 때 전혀 다른 분류 결과가 도출됨을 실험적으로 증명하였고, 이렇듯 모델의 예측 오류를 최대화하는 데이터를 일컬어 적대적 사례(adversarial example)라 정의하였다[6]. 그림 1에서는 적대적 사례의 예시를 보여주며, 좌측의 이미지에 교란값을 추가함으로써 생성된 우측의 이미지들은 모두 타조로 분류된다. 적대적 사례의 존재는 심층신경망 모델의 취약성을 여지없이 보여주며, 적대적 사례를 통해 공격받은 모델은 심지어 높은 확신도(confidence)로 오분류를 수행한다. 이러한 현상은 비단 CNN 기반의 이미지 분류 모델뿐만이 아닌 강화학습이나 RNN(Recurrent Neural Network)을 통해 학습된 모델에서도 발생한다는 사실이 증명되었다[7,8].



〔그림 1〕 적대적 사례 예시[6]. 좌: 원본, 중: 교란값, 우: 적대적 사례 (예측 결과 "타조")

2.2. 적대적 사례의 특징

특정 모델을 공격하기 위해 만들어진 적대적 사례는 다른 신경망 구조로 학습된 모델을 공격할 경우에도 효과적으로 동작한다는 전이성(transferability)이 존재함으로 밝혀짐으로써 적대적 사례에 관한 연구는 심층신경망 학습에 대한 근본적인 문제점으로 여겨지고 인공지능 분야에서의 중요한 주제로 부각되기 시작했다[9].

적대적 사례가 어떻게 존재할 수 있는지에 대한 명확한 이론적 규명은 아직 이루어지지 않았다. 이를 위해서도 다양한 연구들이 수행되었는데, 기존의 연구자들은 단순히 추론 모델이 학습 데이터에 과적합 되어 일반성을 상실했기 때문이라고 유추하였다. Philipp과 Carbonell은 이러한 관점에서 심층신경망의 극단적인 비선형성이 적대적 사례 생성을 가능하게 한다고 보았다[10]. 하지만 Goodfellow등은 오히려 최근의 심층신경망은 모델의 원활한 최적화를 위해 제안된 활성화 함수 및 정규화 방법들로 인해 교란값에 저항하기에는 고차원 특징 공간 대비 충분히 비선형적이지 않으며, 이러한 선형적 특성이 적대적 사례에 취약한 원인이 된다고 주장하였다[11]. Zantedeschi등은 제한 범위가 없는 활성화 함수 이용으로 인해 교란값에 의한 오차가 누적되는 현상으로 설명하였으며[12], Simon-Gabriel등은 적대적 사례가 입력의 차원을 증가시킴으로써 정상적인 모델로서의 동작을 방해한다고 보았다[13]. Moosavi-Dezfooli등은 적대적 사례를 생성하는 교란값이 이미지의 구조보다는 학습 모델 자체에 있다고 보고, 이미지와 무관하게 분류 모델을 공격할 수 있는 보편적(universal) 교란값이 만들어질 수 있음을 실험적으로 증명하였다[14]. Ilyas등은 적대적 사례의 교란값이 어떠한 노이즈가 아닌, 모델 예측시에 분류 성능에 영향을 주는 비강건한(non-robust) 특징의 일종이며 특징 별로 서로 분리(disentangle) 가능하다는 견해를 피력하였다[15].

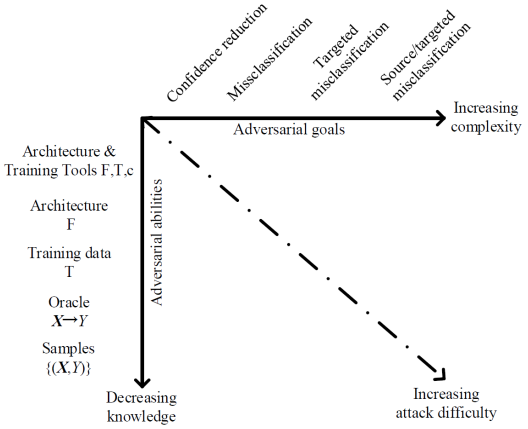
Ⅲ. 적대적 공격과 적대적 사례 생성

3.1. 적대적 공격의 형태

적대적 사례가 심층신경망의 학습 시, 학습 데이터에 주입되어 모델의 정상적 훈련을 방해하는 형태의 공격을 중독(poisoning) 공격이라 하며, 기 학습된 모델의 추론 단계에서 적대적 사례를 통해 오작동을 유도하는 형태의 공격을 회피(evasion) 공격이라 한다[5].

공격자 입장에서는 대상이 되는 심층신경망 모델에 대해 얼마나 많은 사전 정보를 보유하고 있는지에 따라 다음과 같은 세 가지 공격의 경우가 존재한다[16]. 화이트박스(white-box) 공격은 심층신경망의 구조, 하이퍼파라미터 및 학습데이터를 모두 파악하고 있을 때, 그레이박스(grey-box) 공격은 학습된 모델의 파라미터를 제외한 일부 정보를 알고 있을 때의 공격을 의미한다. 상기 두 경우와 같은 경우는 대부분의 심층신경망 모델이 적용된 응용 분야에서 사전 지식의 확보 가능성이 낮으며, 현실적으로 불가능한 상황이다. 세 번째 경우는 블랙박스(black-box) 공격으로 공격자 입장에서는 목표 모델에 대한 지식이 전혀 없으며 입력에 대한 예측 결과만을 관찰할 수 있는 경우이다. 일반적으로 블랙박스 공격은 대상 모델의 예측 결과를 이용해 학습 데이터를 구성한 후, 임의의 심층신경망을 통해 적대적 사례를 생성하여 전이성을 활용함으로써 이루어진다[9].

공격으로써 달성하고자 하는 오분류의 유형 측면에서 공격의 형태를 네 가지로 정의할 수 있다. 공격 대상 모델 자체의 예측 확신도 감소, 분류 결과를 임의의 불특정 클래스로 오분류, 분류 결과를 특정 표적 클래스로 오분류, 그리고 특정 입력에 대해 공격자가 원하는 특정 표적 클래스로의 오분류를 목적으로 수행 할 수 있으며 그림 2에서는 공격자의 사전 정보와 공격 목적을 기준으로 적대적 사례 생성의 난이도를 표현하였다. 공격 대상에 대한 사전 정보가 적고 오분류의 형태가 특정 지어졌을 경우의 공격 난이도가 상승한다[17].



〔그림 2〕 공격자의 사전 지식(세로축)과 공격 목표(가로축)에 따른 공격 난이도[17]

3.2. 기본적인 적대적 사례 생성 기법

3.2.1. L-BFGS[6]

Szegedy등은 효과적인 적대적 사례 $\boldsymbol{x}'$의 생성을 위하여 다음과 같은 최적화 문제를 구성하였다.

$$\min \| \boldsymbol{x}' - \boldsymbol{x} \|_2^2 + c \cdot \ell(F(\boldsymbol{x}'), y_{target}) \quad s.t. \quad x_{\min} \leq \boldsymbol{x}' \leq x_{\max} \tag{1}$$

$\boldsymbol{x}$는 주어진 이미지이며 c는 하이퍼파라미터, $\ell(F(\boldsymbol{x}'), y)$은 적대적 사례가 모델 F에 입력되었을 시의 추론 결과와 목표 레이블 y_{target}간의 교차엔트로피이다. 즉, 이는 적대적 사례 생성을 위한 최소 교란값 $\boldsymbol{r} = \boldsymbol{x}' - \boldsymbol{x}$을 찾기 위함이며, 최적의 해를 찾기 위한 계산 복잡도가 높다는 단점이 있다.

3.2.2. FGSM (Fast Gradient Sign Method)[11]

FGSM은 생성된 적대적 사례의 벤치마크를 위해 이용되는 가장 기본적인 알고리즘이며, 공격 대상 모델 학습과는 반대되는 방향으로 교란값을 추가했을 시 오분류를 유도할 수 있다는 개념이다.

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \cdot sgn(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, F(\boldsymbol{x}))) \tag{2}$$

$J(\boldsymbol{x}, F(\boldsymbol{x}))$는 모델의 학습에 이용된 손실함수이며 $\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, F(\boldsymbol{x}))$는 입력에 대해 계산된 그래디언트이다. ϵ는 교란값의 크기를 조절하는 파라미터이며 $sgn(\cdot)$은 부호 함수이므로 기울기가 증가하는 방향으로 교란값이 생성되며, 결과적으로 간단한 연산만으로도 손실함수 값을 증가시키는 적대적 사례를 생성 가능하다. 특정 표적 클래스 y_{target}로의 오분류를 위한 손실함수 감소 방향의 교란값을 이용한 적대적 사례 공격도 가능하다.

$$\boldsymbol{x}' = \boldsymbol{x} - \epsilon \cdot sgn(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y_{target})) \tag{3}$$

3.2.3. BIM (Basic Iterative Method)[18]

FGSM에서 교란값의 크기 ϵ를 제어하기 어려움에 따라 이의 해결을 위해 $\boldsymbol{x}' \leftarrow \boldsymbol{x}$로 초기화 후, 반복적인 계

산을 통해 조금씩 교란값의 크기를 최적화 하는 방법이다. 매 반복마다 다음과 같이 적대적 사례를 갱신한다.

$$\boldsymbol{x}' \leftarrow clip(\boldsymbol{x}' + \alpha \cdot sgn(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y))) \quad (4)$$

$clip(\cdot)$은 적대적 사례의 값을 제한하는 연산이며 일반적으로 0-255의 값의 범위를 이용한다. BIM은 매 반복마다 픽셀의 값을 1씩 바꿔가며 최적의 적대적 사례를 생성한다.

3.2.4. Iter l.l (Iterative least-likely Class Method)[18]

분류 클래스가 많을 경우, 정답과 유사한 클래스로 오분류되는 것은 성공적인 공격이 아닐 수 있다. 따라서 가장 유사성이 낮은 클래스로 오분류 하도록 적대적 사례를 생성하면 강인한 공격이 될 것이라는 개념에서 착안한 기법이다. BIM과 마찬가지로 반복적 계산을 수행하며, 가장 유사하지 않은 y_{LL} 클래스에 대해 최대가능도를 갖는 방향으로 적대적 사례가 생성된다.

$$\boldsymbol{x}' \leftarrow clip(\boldsymbol{x}' - \alpha \cdot sgn(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y_{LL}))) \quad (5)$$

3.2.5. Carlini-Wagner L_2 Attack[19]

L-BFGS를 개량한 방법으로, 소프트맥스 함수를 통과하기 전 특징의 크기 제어를 통해 교란값을 생성하며 다음과 같은 최적화문제를 구성한다.

$$\min \| \boldsymbol{x}' - \boldsymbol{x} \|_2^2 + c \cdot \ell(\boldsymbol{x}', y) \quad (6)$$

$$\ell(\boldsymbol{x}', y) = \max(\max\{Z_i(\boldsymbol{x}') : i \in Y/\{y\}\} - Z_y(\boldsymbol{x}') + \kappa, 0) \quad (7)$$

$Z(\boldsymbol{x}')$는 소프트맥스 함수를 통과하기 전 마지막 은닉층의 특징이며 κ는 하이퍼파라미터이다. 이는 곧 입력 $\boldsymbol{x}$의 레이블 y가 아닌 임의의 클래스 i로 분류될 수 있도록 특징 공간에서의 L_2 거리를 최대화함을 의미한다. Carlini와 Wagner는 본 적대적 사례 생성 을 이용해 특히 모델 distillation을 통한 적대적 방어기법을 효과적으로 파훼할 수 있음을 보였다.

3.2.6. JSMA (Jacobian-based Saliency Map Attack)[17]

큰 그래디언트 값을 보이는 픽셀이 클래스를 결정하는데 더욱 민감하게 작용한다는 개념을 토대로 만들어진 알고리즘으로, 픽셀 별 미분 $\nabla F(\boldsymbol{x}) = \frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}}$을 통한 그래디언트의 계산을 수행하고 중요맵(saliency map)을 구성한다. 해당 중요맵에서 가장 큰 그래디언트 절대값을 갖는 픽셀에 교란값을 추가한다. 상대적으로 작은 교란값을 통해 특정 표적 클래스로의 오분류를 발생시키는 공격 또한 가능하다는 장점이 있으나 중요맵 구성을 위한 특징별 Jacobian 행렬의 계산이 필수적으로 요구되고, 교란값 생성을 위해 상기 과정을 오분류 발생 시점까지 반복함에 따라 높은 계산복잡도가 요구된다.

3.2.7. DeepFool[20]

DeepFool은 모델이 분류를 위해 선형 초평면 결정경계를 갖는다고 가정한다. 특징 공간에서 입력 $\boldsymbol{x}$와 가장 가까운 결정경계의 거리를 계산한 후, 해당 거리만큼의 크기를 갖는 교란값을 생성한다는 것이 기본적인 알고리즘이다. 이를 위해 다음과 같은 문제를 정의하였다.

$$\begin{aligned} &\text{argmin}_{\boldsymbol{r}_i} \| \boldsymbol{r}_i \|_2 \\ &s.t. \quad g(\boldsymbol{x}_i) + \nabla g(\boldsymbol{x}_i)^T \boldsymbol{r}_i = 0 \end{aligned} \quad (8)$$

$g(\cdot)$는 분류 함수이다. 본 알고리즘에서 교란값 $\boldsymbol{r}$은 i번째 마다 반복적으로 갱신되며, $g(\boldsymbol{x}_{i+1})$의 부호가 바뀔 때, 교란값의 크기가 최소화 되었다 판단하고 반복이 종료된다. 가정이 선형 결정경계이므로, 매 반복 i번째 마다 달라지는 위치인 $\boldsymbol{x}_i$에 대한 미분을 통해 $g(\cdot)$가 선형으로 근사화되며, 그 거리를 계산해 $\boldsymbol{r}_i$를 계산하고 $\boldsymbol{x}_i$에 더함으로써 위치 역시 갱신하는 방식으로 적대적 사례를 생성한다.

Ⅳ. 최근 적대적 공격

4.1. 디지털 공격과 물리적 공격

최근 적대적 공격은 이미지의 직접 입력 가능 여부에

따라 다음과 같은 두 경우로 나눌 수 있다[28]: 1) 디지털 세팅: 공격자가 기 학습된 심층신경망 분류 모델에 직접 디지털 데이터 형태의 입력을 할 수 있는 경우, 2) 실세계(physical-world) 세팅: 심층신경망 분류 모델에 직접 입력 정보를 전달할 수 없고, 시스템이 제공하는 이미지 센서(카메라 촬영)를 통해서만 이입력이 가능한 경우.

현재 대부분의 적대적 사례 생성 기법은 디지털 세팅 기반의 공격을 염두에 두고 있으며, 이는 생성된 적대적 사례를 직접 공격 대상에 입력할 수 있음을 뜻한다. 하지만 실제 응용 시스템들은 심층신경망으로의 직접 입력을 불허하며, 외부에 장착된 시스템의 카메라를 통해서만 입력을 허용하는 경우가 다수이다. 발전된 기계학습 및 이미지 처리 기법 기반으로 디지털 세팅에 대한 적대적 사례 생성 연구가 지속적으로 진행됨과 동시에, 실세계 세팅을 고려한 적대적 사례의 생성에 대한 관심 역시 날로 증가하고 있다.

4.2. 최근 적대적 사례 생성 기법

4.2.1. SemanticAdv (Semantic Adversarial Examples)[21]

기존의 적대적 사례 생성 기법들은 최대한 사람의 눈에 띄지 않는 노이즈 형태의 작은 교란값을 발생시켜 기 학습된 모델을 오작동 하도록 설계됨을 목표로 하였다면, 오히려 본 방법은 교란값을 인간의 눈으로 확인 가능하고 의미가 있는 형태로 생성함으로써, 육안으로는 입력 이미지의 클래스를 분류하는데 이상을 발견할 수 없으나 심층신경망 모델은 오분류를 유도하도록 적대적 사례를 만드는 방법이다.

Hosseini와 Poovendran은 이미지를 HSV 색공간으로 변환하여 각 채널 별 약간의 이동(shifting)을 통해 교란값이 생성될 수 있음을 보였으며 이를 위해 다음과 같은 최적화 문제를 정의하였다.

$$\min |\boldsymbol{r}| \quad s.t. \begin{cases} \boldsymbol{x}_H' = (\boldsymbol{x}_H + \boldsymbol{r}_H) \bmod 1 \\ \boldsymbol{x}_S' = clip(\boldsymbol{x}_S + \boldsymbol{r}_S, 0, 1) \\ \boldsymbol{x}_V' = \boldsymbol{x}_V \end{cases} \tag{9}$$

본 알고리즘은 이미지의 밝기를 제외한 색조와 채도 성분을 조금씩 변화시켜가며 공격 대상 모델이 오분류

〔그림 3〕 SemanticAdv. 좌: 원본, 우: 컬러 이동을 통한 적대적 사례[21]

를 일으킬 때까지 반복함으로써 색조와 채도 성분을 이동시키는 최적의 교란값 $\boldsymbol{r}_H$와 $\boldsymbol{r}_S$을 생성한다. 그림 3에서는 색조와 채도 변화를 통해 오분류되는 예제를 보여준다. 생성된 적대적 사례는 컬러성분만 변화되었을 뿐 육안으로 확인하였을 시 의미적으로 분류에 문제가 없으나, 심층신경망의 오동작을 효과적으로 유발한다.

4.2.2. Edgefool[22]

기존 적대적 사례들은 고주파 성분을 주로 가지는 노이즈 형태의 교란값을 생성함에 따라, 공격을 성공률을 높이기 위해 교란값의 크기를 키웠을 경우에는 육안으로의 확인이 용이하다는 단점이 존재한다. Edgefool은 일반적으로 이미지 향상을 위한 기존의 필터링 기법들이 엣지나 공간적 고주파 성분의 에너지를 증가시킨다는 성질에 기반하여 해당 영역에 집중적으로 교란값을 추가함으로써 오히려 높은 품질의 이미지로 보이도록 적대적 사례를 생성하는 기법이다. 입력 이미지의 색공

〔그림 4〕 좌: 원본, 중: SemanticAdv, 우: Edgefool〔22〕

간을 Lab로 변환한 후, 색차 성분은 유지하되 밝기 성분의 디테일을 향상시키면서 교란값 해당 부분에 교란값을 추가하는 모든 층이 콘볼루션 연산으로 구성된 네트워크를 학습시켜 이용하였으며, 그림 4에서 확인할 수 있듯 이미지의 선명도 향상 기법을 적용한 결과물과 같은 형태의 적대적 사례를 생성하였다.

4.2.3. ColorFool[23]

SemanticAdv의 경우 비제한적(unrestricted)으로 교란값을 생성시킬 수 있다는 장점이 있으나, 교란값의 크기가 커지면서 비자연스러운(unnatural) 적대적 사례가 생성될 수 있다는 단점이 있다. ColorFool은 이를 해결하고자 이미지의 영역을 의미론적 영역을 K개로 구분하여 사람, 하늘, 식물, 그리고 물 영역은 컬러에 대한 중요도와 인간의 주목도가 높은 영역 $\boldsymbol{s}$로 정의하였고, 그 외의 영역은 컬러가 상대적으로 크게 이동되어도 괜찮은 영역 $\bar{\boldsymbol{s}}$라 규정하였다. 이에 Lab 색공간에서 색차 성분에 대해 다음과 같은 연산을 통하여 적대적 사례를 생성하였다.

$$\boldsymbol{x}' = Q\left(\gamma^{-1}\left(\begin{matrix}\sum_{k=1}^{S}\left(\gamma(\boldsymbol{s}_k)+\alpha\left[0, \boldsymbol{r}_k^a, \boldsymbol{r}_k^b\right]^T\right) \\ +\sum_{k=1}^{S}\left(\gamma(\bar{\boldsymbol{s}}_k)+\alpha\left[0, \bar{\boldsymbol{r}}_k^a, \bar{\boldsymbol{r}}_k^b\right]^T\right)\end{matrix}\right)\right) \tag{10}$$

여기서 $\gamma(\bullet)$는 RGB에서 Lab로 색공간은 변환하는 연산을 의미하며, $Q(\bullet)$는 교란값 추가를 통해 생성된 이미지의 다이나믹레인지를 제어하는 양자화 함수를 의미한다. 그림 5에서는 ColorFool을 이용해 생성된 적대적 사례의 예제를 나타내며 객체는 그대로 두고 배경의 색 변화만으로 육안으로 보았을 때에는 의미적으로 이상을 인지할 수 없으나 심층신경망 분류 모델에 입력되었을 시 오류를 발생시킬 수 있음을 확인 가능하다 (원본 클래스: 정장, 오류 클래스: 라켓).

〔그림 5〕 좌: 원본, 우: ColorFool〔23〕

4.2.4. SPA (Structure Preserving Attack)[24]

SPA는 인간의 시각 체계가 구조적 정보에 민감하게 반응한다는 특성을 반영하여 기존 랜덤하게 발생하는 노이즈 형태의 교란값을 제어함으로써 이미지의 구조적 패턴과 일치하는 형태의 교란값을 생성할 수 있음을 보였다. 이를 통해, 이미지에서의 구조적 특성을 유지하여 육안으로는 눈에 띄지 않으나 큰 값을 갖는 교란값을 발생시켰고, 또한 해당 기술로 생성된 적대적 사례들은 높은 전이성을 가짐을 확인하였다.

알고리즘의 핵심은 기존 L_p-norm 기반의 적대적 사례 생성 기법들이 각 픽셀별 교란값의 발생을 상호 독립적으로 수행하였음에 문제가 있다 판단하고, 같은 구조 내 픽셀들은 서로 비슷한 교란값을 갖도록 유도한다는데 있다. 이를 위해 메타 교란값의 개념을 도입하여 구역별로 교란값을 발생시킨 뒤, 이미지의 분할(segmentation) 정보를 통해 반복적으로 교란값을 클러스터링 해가며 결정함으로써 최종적으로는 주변 픽셀과 유사한 교란값을 생성할 수 있도록 한다.

〔그림 6〕 좌: 원본, 중: SPA의 교란값, 우: SPA〔24〕

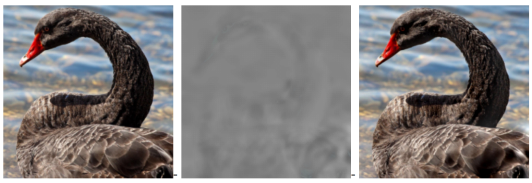
4.2.5. Shadow Attack[25]

Ghiasi 등은 강력한 적대적 사례의 방어 기법 중 하나인 인증된(certified) 분류기의 공격을 위해 비가시성을 보장하면서도 큰 교란값을 발생시킬 수 있는 적대적 사례 생성 기법인 Shadow attack을 제안하였다. 인증된 분류기란, 입력 이미지가 주어졌을 때 L_p 범위 내 적대

적 사례가 수학적으로 생성될 수 없도록 강건함(robustness)을 보장하는 방어 기법이다 (예: random smoothing 기법). Shadow attack은 인증된 분류기가 매우 큰 교란값을 갖게 할 수 있다는 한계점을 지적하고 다음과 같은 목적함수를 통해 최적화된 적대적 사례를 생성하였다.

$$\max_{\boldsymbol{r}} L(\theta, \boldsymbol{x}+\boldsymbol{r}) - \lambda_c C(\boldsymbol{r}) - \lambda_{tv} TV(\boldsymbol{r}) - \lambda_s Dissim(\boldsymbol{r}) \quad (11)$$

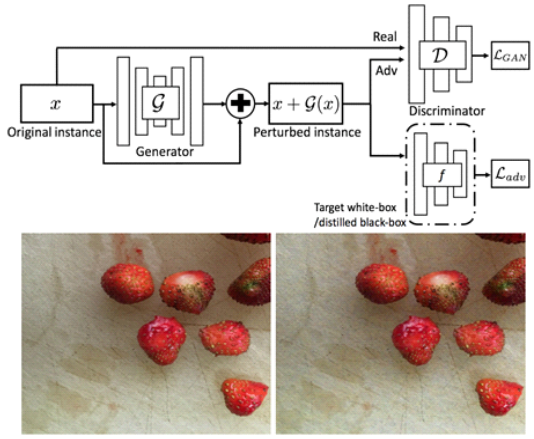
$L(\bullet)$은 손실함수이며 $C(\bullet)$는 교란값의 전체적인 스케일을 최대화 하기 위한 항, $TV(\bullet)$는 총 변이(total variation)로써 인접 픽셀과의 유사도를 최대화하기 위한 항, 그리고 $Dissim(\bullet)$은 이미지의 각 채널별 유사도를 최대화하기 위한 항이다. $\lambda_c, \lambda_{tv}, \lambda_s$는 각 항의 페널티를 의미하는 하이퍼파라미터이며, (11)을 통해 그림 7과 같이 그림자 형태의 값이 큰 교란값을 생성할 수 있었으며, 실제 사람의 육안으로는 확인이 어려운 적대적 사례를 이용해 인증된 분류기를 효과적으로 공격할 수 있었다.



〔그림 7〕 좌: 원본, 중: Shadow Attack의 교란값, 우: Shadow Attack을 통한 적대적 사례〔25〕

4.2.6. AdvGAN (Adversarial GAN)[26]

노이즈제거, 깊이값 계산, inpainting, 이미지 화질 향상, 초해상도 등 수 많은 이미지 처리 연구는 이미 GAN(generative adversarial network)을 이용하여 원하는 목적의 결과물을 생성하는 형태로 진행되고 있으며, Xiao등은 이러한 맥락에서 GAN을 이용해 적대적 사례를 생성하는 기법을 소개하였다. 그림 8에서는 AdvGAN의 전체적인 구조를 나타내었으며, 생성망 $g(\bullet)$가 판별망 $D(\bullet)$로 정의된 적대적 손실함수를 최소화하고, 공격 대상이 되는 분류기 $f(\bullet)$의 오분류율을 최대화 하는 방향으로 학습됨으로써 최적의 교란값을 생성한다. 클라우드 소싱을 이용한 유저스터디를



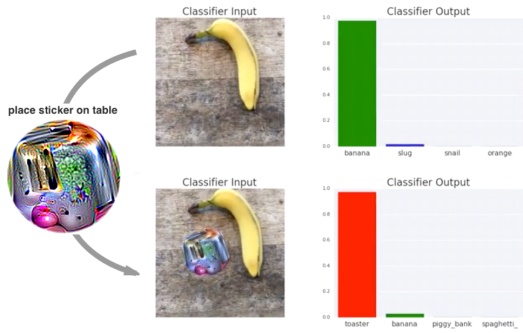
〔그림 8〕 AdvGAN의 구조와 생성된 적대적 사례. 분류 결과 좌: "딸기", 우: "장난감 개"〔26〕

통해 AdvGAN을 통해 생성된 적대적 사례들이 실제 인지적으로도 사람들의 눈에 확인되지 않음을 보였고, 특히 블랙박스로 세팅된 $f(\bullet)$ 공격 시에도 효과적인 적대적 사례를 생성함을 확인하였다.

4.2.7. AdvPatch (Adversarial Patch)[27]

대표적인 실세계 세팅에서의 공격 방법이며, Brown 등은 카메라로 촬영할 실제 객체 주변에 단순히 붙이기만 함으로써 보편적(universal)으로 동작하는 패치 형태의 적대적 사례가 존재함을 보였고, 이를 생성하여 활용할 수 있도록 하였다.

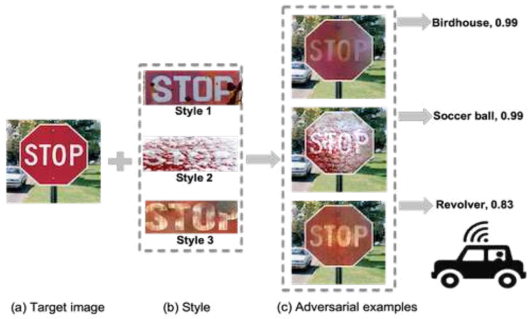
제작된 적대적 패치는 그림 9와 같으며, 패치 부착의 위치, 패치의 회전 및 크기에 관계없이 오분류를 일으킬 수 있도록 제작되었다.



〔그림 9〕 AdvPatch: 실세계 세팅의 적대적 공격〔27〕

4.2.8. AdvCam (Adversarial Camouflage)[28]

Duan등은 디지털 및 실세계 세팅에 모두 적용 가능한 형태의 적대적 사례를 생성하고자 하였다. 디지털 세팅 가정으로 제안된 대부분의 적대적 사례 생성 기술들이 인간으로 하여금 교란값을 인지 불가능하도록 목적함수를 두어 생성하는데, 이러한 수준의 작은 교란값으로는 실세계 세팅에서 현실적으로 동작이 어려움을 인정하였다. 이에 그림 9와 같이 디지털 이미지에 실세계에서 발생 가능한 특정 스타일 형태의 교란값 패턴을 생성하였고, 해당 패턴을 실세계 세팅에 적용하였을 때에도 높은 적대적 공격 성공률을 보임을 확인하였다.



〔그림 10〕 디지털 및 실세계 세팅에 모두 적용 가능한 AdvCam〔28〕

V. 결 론

본 논문에서는 기존 적대적 사례의 생성 기법과 쟁점에 대해 살펴보고 근래의 인공지능 및 컴퓨터비전 분야 주요 학회에서 소개되는 이미지 기반 적대적 사례 생성 기술 동향에 관해 분석하였다.

앞으로의 연구는 이론을 통한 적대적 사례의 생성보다는 다양한 도메인에서 실제 상황(in the wild)으로의 활용 가능성 모색을 위한 실세계 세팅에서의 연구가 활발히 진행될 것으로 예상된다.

적대적 사례를 통한 공격에 대항하고자 이미 심도 깊은 수준의 적대적 방어 기술들이 소개되고 있으며, 모델의 강건함은 지속적으로 높아지고 있다. 이에, 특정 방어에 특화된 적대적 사례 생성 기술이 지속적으로 소개될 것으로 보이며, 또한 심층신경망의 설명 가능한 동작에 관한 원론적인 논의도 꾸준히 이루어질 것으로 생각된다.

참 고 문 헌

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Neural Information Processing Systems (NIPS)*, 2012.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] 김휘영, 정대철, 최병욱, “딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격,” *대한영상의학회지*, vol. 80, no. 2, pp.259-273, 2019.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xia, and D. Song, “Robust physical-world attacks on deep learning visual classification,” *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] J. Zhang, and C. Li, “Adversarial examples: opportunities and challenges,” *IEEE Trans. Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2578-2593, 2020.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *Int’l Conf. Learning Representation (ICLR)*, 2014.

[7] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, “Adversarial attacks on neural network policies,” *Int’l Conf. Learning Representation (ICLR)*, 2017.

[8] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, “Deep text classification can be fooled,” *Int’l Joint Conf. Artificial Intelligent (IJCAI)*, 2018.

[9] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” *ACM Asia Conf. Comput. Comm. Security*, 2017.

[10] G. Philipp, and J. G. Carbonell, “The nonlinearity coefficient – predicting generalization in deep neural network,” *arXiv:1806.00179,* 2018.

[11] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *Int’l Conf. Learning Representation*

(ICLR), 2015.

[12] V. Zantedeschi, M. Nicolae, and A. Rawat, “Efficient defenses against adversarial attacks,” *ACM Workshop on Artificial Intelligence and Security (AISec),* 2017.

[13] C-J. Simon-Gabriel, Y. Oliver, L. Bottou, B. Scholkopf, and D. Lopez-Paz, “First-order adversarial vulnerability of neural networks and input dimension,” *Int’l Conf. Machine Learning (ICML)*, 2019.

[14] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” *Neural Information Processing Systems (NeurIPS)*, 2019.

[16] B. Biggio, and F. Roli, “Wild patterns: ten years after th rise of adversarial machine learning,” *Pattern Recog.*, vol. 84, pp. 317-331, 2018.

[17] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016.

[18] A. Kurakin and I. Goodfellow, “Adversarial examples in the physical world,” *Int‘l Conf. Learning Representations (ICLR)*, 2017.

[19] N. Carlini and D. Wagner, “Defensive Distillation is Not Robust to Adversarial Examples,” *arXiv:1607.04311*, 2016.

[20] S.M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: a simple and accurate method to fool deep neural networks,” *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] H. Hosseini, and R. Poovendran, “Semantic adversarial examples,” *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2018.

[22] A. S. Shamsabadi, C. Oh, and A. Cavallaro, “Edgefool: an adversarial image enhancement filter,” *Int’l Conf. Acoustics, Speech and Signal Process. (ICASSP)*, 2020.

[23] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, “ColorFool: semantic adversarial colorization,” *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] D. Peng, Z. Zheng, L. Luo, and X. Zhang, “Structure matter: towrads generating trasnferable adversarial images,” *European Conf. Artificial Intelligent (ECAI)*, 2020.

[25] A. Ghiasi, A. Shafahi, and T. Goldstein, “Breaking certified defenses: semantic adversarial examples with spoofed robustness certificates,” *Int’l Conf. Learning Representation (ICLR)*, 2020.

[26] C. Xiao, B. Li, W. He, M. Liu, and D. Song, “Generating adversarial examples with adversarial networks,” *Int’l Joint Conf. Artificial Intelligent (IJCAI)*, 2018.

[27] T. Brown, D. Mane, A. Roy, M. Adabi, and J. Gilmer, “Adversarial patch,” *Neural Information Processing Systems (NIPS) Workshop,* 2017.

[28] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yun, “Adversarial camouflage: hiding physical-world attacks with natural styeles,” *Computer Vision and Pattern Recognition (CVPR)*, 2020.

〈저자소개〉

오 희 석 (Heeseok Oh)

2017년 2월 : 연세대학교 전기전자공학과 박사
2017년 2월~2017년 8월 : 삼성전자 DMC연구소 책임연구원
2017년 9월~2020년 2월 : 한국전자통신연구원 선임연구원
2020년 3월~현재 : 한성대학교 IT융합공학부 조교수
<관심분야> 영상처리, 컴퓨터비전, 혼합현실, 심층생성모델 등