다양한 지각 불가능 영역 탐색 기법을 통합한 적대적 예제 생성 기법 연구

박윤수*, 이상호*, 조영호(교신저자)**
*국방대학교 국방관리대학원 사이버컴퓨터공학과 석사과정
**국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과 교수
e-mail:younghocho@korea.kr

A Study of Integrating Multiple Imperceptible Region Detection Techniques for Adversarial Example Generation Method

Yoonsoo Park*, Sangho Lee*, Youngho Cho**
*Master's Course, Dept. of Cyber Security and Computer Engineering, Korea National Defense University
**Professor, Dept. of Cyber Security and Computer Engineering, Korea National Defense University

요 약

딥러닝(Deep Learning) 기술은 많은 분야에서 높은 성능을 보이고 있지만 입력 데이터에 의도적으로 노이즈를 추가하는 모델의 오분류를 유도하는 적대적 예제 공격(Adversarial Example Attack)에 취약한 것으로 알려져 있다. 그러나 원본 데이터에 추가된 노이즈가 적대적 예제의 지각적 품질을 일정 수준 이상으로 훼손하면 인간의 시각체계와 인공지능(AI) 기반 탐지 모델에 의해 발각될 수 있다. 이러한 단점을 극복하기 위해 지각적 품질을 고려한 적대적 예제 기법이 제안되었다. 대표적으로 복잡도 지각 모델(Complexity Perception Model)과 분산 맵(Variance Map) 등이 있으며 최근에는 두 기법들을 통합한 C&V 기법이 제안되었다. 특히, C&V 기법은 지각 불가능 영역을 확장하여 공격 성공률을 높이고 지각적 품질을 유지하였으나 픽셀을 사전에 지정한 비율에 따라 무작위로 선정하여 공격 자원이 비효율적으로 사용하였다. 또한, 지각적 품질을 고려한 적대적 예제 기법들이 다수 존재하고 있으며 이를 더 통합하여 최적화 한다면 보다 성공적인 적대적 예제를 생성할 수있다. 따라서, 본 논문에서는 기존 기법들에 비해 보다 높은 공격성공률을 달성하면서도 지각적 품질을 유지할 수 있는 적대적 예제 생성 기법을 제안한다. 구체적으로, 여러 지각불가능 영역 탐지 기법들을 활용하여 확장된 지각 불가능 영역을 구축하고 Gradient Magnitude를 활용하여 공격 성공률이 높은 픽셀을 선별하고 점수화(Scoring)한 후 높은 점수를 가진 픽셀에 우선적으로 공격기법을 적용하여 지각적 품질과 공격 성공률을 높일 수 있음을 확인하였다.

1. 서론

딥러닝(Deep Learning) 기술은 이미지와 음성, 언어 등 여러 분야에서 탁월한 성능을 보이고 있으나 이러한 모델은 입력 이미지에 미세한 노이즈를 추가하는 방식으로도 쉽게 오작동을 유도할 수 있는 적대적 예제(Adversarial Example) 공격에 취약하다는 점이 알려지면서 이에 대응하는 많은 연구가 진행되고 있다[1, 2].

일반적으로 이미지를 입력으로 하는 적대적 예제 공격은 출력 이미지가 탐지되지 않도록 하기 위해 원본 이미지와 최대한 유사하게 이미지를 생성한다. 그러나 정상적인 이미지에 적대적 예제를 생성하기 위해 노이즈를 추가하면 생성된 적대적 예제의 시각적 품질이 심각히 저하될 수 있으며, 그에 따라 시각적 품질이 낮은 적대적 예제 이미지는 쉽게 탐지 및 제거될 수 있다[2].

이러한 문제를 해결하기 위해 주어진 이미지 내에서 약간의 수정이 될 경우에도 인간의 눈에는 쉽게 인지되지 못하는 픽셀들의 집합을 뜻하는 지각 불가능 영역(Imperceptible Region)을 탐색하는 기술이 제안되었다[3, 4, 5]. 대표적으로 복잡도 지각 모델(Complexity Perception Model, CPM)[3], 분산 맵(Variance Map, VM)[4], 그리고 이 두 가지 기법들을 통합하여 지각 불가능 영역을 확장한 방식인 AEGM-C&V[5]가 있다. 그러나 C&V 기법에서 섭동(Perturbation)을 추가할 픽셀들을 다양한 비율로 랜덤하게 선정하여 성능을 평가한 결과를 실험적으로 제시한 반면에, 최적의 공격성공률과 지각적 품질을 달성하기 위한 픽셀 조합을 어떻게 찾을 지에 대한 연구는 미흡하였다[5].

따라서, 본 논문에서는 기존 기법의 제한사항을 극복하기 위해 다수의 지각 불가능 영역 탐색 기법들을 통합하여 지각 불가능 픽셀영역을 확장하고, 최적의 공격 성공률과 지각적 품질을 고려한 적대적 예제 생성을 위한 스코어 기반(Scoring-based) 섭동 삽입 기법을 제안한다.

이후 논문의 구성은 다음과 같다. 2장에서는 배경지식과 관련 연구를 소개하고, 3장에서는 연구문제를 기술한 후 해결 아이디어를 제안한다. 4장에서는 초도실험을 통해 제안 아이디어의 유

효성을 확인하고, 끝으로 5장에서는 결론과 향후 연구방향을 기술한다.

2. 배경지식 및 관련 연구

2.1 적대적 예제 기법

딥러닝 모델은 높은 정확도에도 불구하고 입력에 대한 노이즈에 취약한 구조를 가지고 있으며, 이러한 취약점을 이용한 공격 방식이 적대적 예제이다[1]. Szegedy 등[1]은 입력 이미지에 인간이 인지하지 못할 정도의 작은 변화를 주었을 때 모델이 오분류를 일으킬 수 있다는 사실을 발견했다. 또한, Goodfellow 등[6]은 모델의 손실함수가 증가하는 방향으로 일정 크기의 노이즈(ϵ)를 추가함으로써 빠르고 효율적인 FGSM(Fast gradient Sign Method)을 (1)과 같이 제안했다.

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y)) \quad (1)$$

이때, x는 원본 입력 이미지, x_{adv}는 생성된 적대적 이미지, ϵ은 공격 강도, $\nabla_x J(\theta, x, y)$는 입력 x에 대한 손실 함수의 기울기, $sign()$은 실수 입력의 부호를 출력하는 함수로 손실 함수의 기울기 방향을 의미한다. 즉, 손실 함수 J를 통해 모델이 정답과 얼마나 다른지를 평가하고 손실을 가장 빠르게 증가시키는 방향, 즉 기울기 방향으로 노이즈를 삽입한다. 여기서 $sign$은 방향을 통해 각 픽셀이 ±변화만 부여함으로써 공격자가 허용한 동일한 크기(ϵ)만큼 변화하도록 한다.

2.2 지각적 품질 지표(Perceptual Quality Metrics)

적대적 예제는 성공적으로 오분류를 유도하더라도 지각적으로 부자연스럽거나 눈에 띈다면 실제 공격에서는 탐지될 가능성이 높아진다. 즉, 적대적 예제가 모델을 속이더라도 인간에게 노출될 경우 탐지되거나 무력화될 수 있기 때문에 지각적 품질은 적대적 공격의 실용성과 은닉성을 동시에 평가하는 핵심 지표로 사용된다. 지각적 품질을 평가하기 위해 다양한 지표들이 있으며 대표적으로 다음과 같다.

2.2.1 PSNR(Peak Signal to Noise Ratio)

PSNR[3]은 원본 이미지와 적대적 예제간의 평균 제곱 오차(Mean Squared Error, MSE)를 기반으로 유사성을 측정하는 지표로써 다음과 같은 수식 (2)로 계산된다. 이때, MAX는 픽셀의 최대값이며, MSE는 두 이미지 간의 평균 제곱 오차이다. PSNR의 값이 높을수록 원본 이미지와 유사함을 의미하며 일반적으로 30dB 이상이면 인간의 시각 체계(Human Visual System, HVS)로 차이를 인식하기 어렵다.

$$PSNR = 10 \cdot log_{10}(\frac{MAX^2}{MSE}) \quad (2)$$

2.2.2 SSIM(Structural Similarity Index Measure)

SSIM[3]은 인간의 시각 체계가 민감하게 반응하는 밝기, 대비, 구조의 요소를 반영하여 두 이미지 간의 시각적 유사성을 평가하는 지표이다. SSIM의 수식은 (3)과 같다.

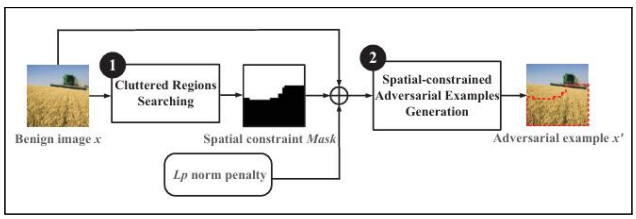
$$SSIM(x,y) = \frac{(2\mu_x\mu_u + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

이때, μ_x, μ_y는 이미지 x와 y의 평균 픽셀 값, σ_x, σ_y은 이미지 x와 y의 픽셀 분산 값, σ_{xy}는 x와 y간의 공분산, C_1, C_2는 상수를 나타낸다. SSIM은 두 이미지 간의 유사성을 평가하는 지표로써 0에서 1 사이의 결과를 갖으며, 1에 가까울수록 원본 이미지와 더 유사함을 의미한다.

2.3 지각 불가능 영역(Imperceptible Regions)

지각 불가능 영역이란 인간의 시각 체계가 민감하게 반응하지 않거나 주의 깊게 보지 않는 이미지의 특정 영역을 의미한다[5]. 이러한 영역에만 노이즈를 삽입할 경우, 적대적 예제는 시각적으로 자연스러움을 유지하면서 딥러닝 모델을 효과적으로 공격할 수 있으며 이는 탐지 가능성을 현저히 낮추는 데 기여한다.

2.3.1 복잡도 지각 모델(CPM)



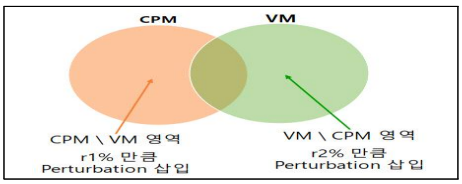
[그림 1] CPM 흐름도[3]

CPM은 Wang 등[3]에 의해 제안된 모델로, HVS는 복잡도가 높은 시각 정보에 대해 민감도가 낮다는 특성에 기반한다. CPM은 저수준 이미지 특성을 선형적으로 결합하여 각 패치의 복잡도를 수치화한다. 그림 1에서와 같이, 복잡도 기반 탐색(Complexity Estimation)을 통해 지각적으로 민감하지 않은 영역, 즉 복잡한 영역(Cluttered region)을 찾아낸다. 결과적으로 Spatial Constraint Mask가 생성되고 이는 어디에 공격을 삽입할 수 있는지를 나타내는 마스크이며 CPM에서 생성한 마스크를 기반으로 공격 영역을 제한하면서 적대적 예제(x')를 생성한다.

2.3.2 분산맵(VM)

VM은 Croce 등[4]에 의해 제안되었으며 HVS는 이웃 픽셀 간의 분산(variance)이 낮은 영역에 추가된 변형(perturbation)에 더 민감하며 분산이 높은 영역에서는 이러한 변형을 덜 인지한다는 특성에 기반한다. 즉, 고분산(high variance) 영역은 변화가 많아 인간의 주의가 덜 집중되며 그 영역에 노이즈를 삽입하면 시각적으로 인지하기 어렵다.

2.3.3 AEGM-C&V



[그림 2] AEGM-C&V: CPM과 VM 결합

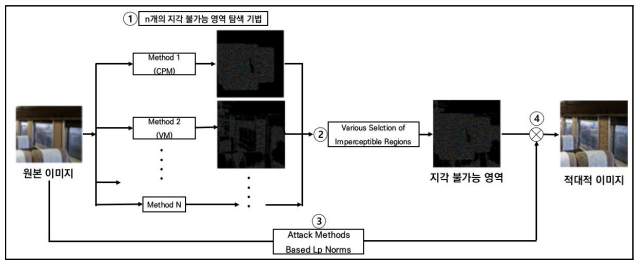
AEGM-C&V(이하 C&V)는 이륜건 등[5]에 의해 제안되었으며, 이미지로부터 확장된 지각 불가능 영역을 찾기 위해 CPM기법과 VM기법을 결합하였다. 이미지 내 CPM과 VM이 공통으로 해당되는 교집합 영역에는 모든 픽셀에 공격기법을 적용하며 차집합 영역인 CPM \ VM과 VM \ CPM에는 사전에 지정한 비율 r1, r2에 따라 공격기법을 적용한다. 예를 들어, r1:r2 = 0.2:0.4 일 때, CPM-VM 영역에서 20%의 픽셀에 무작위로 공격기법을 적용하며 VM-CPM 영역에서 40%의 픽셀에 무작위로 공격기법을 적용한다. 그 결과, CPM, VM을 단독으로 활용했을 때 보다 더 향상된 공격 성능을 보였다.

3. 제안기법

3.1 문제 기술

기존의 C&V 기법은 이미지의 지각 불가능 영역을 확장하여 공격 기법을 적용하는 영역이 증가하였기 때문에 기존 CPM과 VM 기법과 비교하여 SSIM과 PSNR은 유지한 채 공격 성공률(ASR)은 증가시킬 수 있었다. 그러나 C&V 기법은 차집합 영역에서 비율에 따라 픽셀을 무작위로 선정하여 공격 기법을 적용한다. 따라서 얼마만큼의 픽셀에 공격 기법을 적용해야 최적의 조건을 달성하는지를 하나씩 측정해야 하며 최적의 비율을 산정하는데 시간이 소요되는 단점이 있다. 또한 지각 불가능 영역 탐색 기법은 소개한 기법보다 더 많이 존재하므로 3개 이상의 다수 기법을 통합한 추가 확장 가능성이 있다.

3.2 제안기법 설명



[그림 3] 제안 기법 흐름도

기존의 C&V 기반 적대적 예제 생성 기법에서는 차집합 영역에서 공격 픽셀을 무작위로 선택하여 비율에 따라 공격기법 I 을 적용하였다. 이러한 방법은 공격 기대 효과가 낮은 픽셀에 공격 기법이 적용될 수 있어 효과적이지 못할 수 있다. 본 연구에서는 이러한 한계를 극복하기 위해 단계별로 제안 기법을 소개한다 (그림 3 참조).

• **1단계 (지각 불가능 영역 탐색 단계)**: n개의 지각 불가능 영역 탐색 기법들을 활용하여 지각 불가능 영역을 찾는다. CPM과 VM 외에도, 색공간을 활용한 YUV 기법[7], 주파수 도메인을 활용한 고주파 교란 기법[8] 등 기법들이 있다.

• **2단계 (지각 불가능 영역 통합 단계)**: 1단계에서 각기법들이 찾은 지각 불가능 영역에 속하는 픽셀들 중 중복된 픽셀들은 제외하여 통합하여 식별한다.

• **3단계 (픽셀별 점수화 및 순위부여 단계)**: 각 픽셀에 대해, 공격성공률과 지각적 품질 향상에 대한 영향도를 점수화하고 순위를 부여한다. C&V 기법과는 다르게 제안 기법은 차집합 영역에서는 픽셀의 민감도를 측정하여 높은 민감도를 가진 픽셀은 공격 성공률이 높을 것이라는 가정하에 순위를 매긴다. 이때 픽셀의 점수를 측정하여 순위를 매기게 되는데 'Gradient Magnitude(기울기 크기)'를 도입하여 픽셀의 점수를 측정하고 순위를 매긴다. Gradient Magnitude는 입력 이미지의 각 픽셀에 대해 손실 함수의 변화율을 벡터 형태로 계산한 수, 그 크기를 측정한 값으로 정의된다. 즉, 해당 픽셀이 손실 함수에 얼마나 민감한지를 수치적으로 나타내는 지표이며 수식 (4)은 다음과 같다.

$$\| \nabla_x J(\theta,x,y) \|_2 = \sqrt{(\frac{\partial J}{\partial x_r})^2 + (\frac{\partial J}{\partial x_g})^2 + (\frac{\partial J}{\partial x_b})^2} \quad (4)$$

이때, $J(\theta,x,y)$는 입력 x와 정답 레이블 y, 파라미터 θ에 대한 손실 함수를 나타내며, $\nabla_x J$ 는 입력 x에 대한 손실 함수의 기울기(gradient)를 나타낸다. 즉, Gradient Magnitude가 클수록 해당 픽셀은 모델의 분류 결과에 큰 영향을 미치며, 이러한 픽셀에 공격 기법을 적용한다면 공격 성공률(ASR)이 높아질 가능성이 크다.

• **4단계 (섭동 삽입 단계)**: 상위 점수를 갖는 픽셀부터 공격자가 허용한 픽셀수만큼 선정하여 섭동(Perturbation)을 삽입함으로써 적대적 예제를 생성한다. 즉, 모든 픽셀을 대상으로 공격하지 않고, 필요한 만큼 섭동 추가 대상 픽셀을 선정하여 공격함으로써 지각적 품질을 유지할 수 있다.

4. 초도실험

4.1 실험목적 및 방법

초도실험의 목적은 제안 기법의 실현가능성과 성능을 확인하는 것이다. 구체적으로, 제안기법의 세 번째 단계(픽셀별 점수화 및 순위부여) 적용을 통한 성능 향상 효과를 확인하고 분석하는 것이다. 탐색된 지각 불가능 영역 픽셀 중 상위 50% 픽셀을 대상으로 FGSM을 활용하여 섭동을 추가하는 공격을 수행하였다.

실험 환경 구성으로 공격 대상 모델은 Inception V3, 데이터셋은 NIPS2017 이미지 1,000장을 사용하였다.

4.2 초도실험 결과 및 분석

우선, 표 1은 4개의 비교 대상 기법들(CPM, VM, C&V, 제안기법)이 가장 좋은 최대 성능을 내는 결과값을 비교하여 나타낸 것이다. 즉, C&V는 가능한 비율 조합 중 최상위 성능을 내는 것으로 비교하였으므로 제안기법에 매우 불리한 조건임을

의미한다.

[표 1] 비교 실험결과(불공정 비교, 최대 효과 기준)

기법	ASR	SSIM	PSNR
CPM 단일	64.3%	0.970	44.77
VM 단일	74.3%	0.957	43.94
C&V(0.2:0.4)	74.0%	0.961	44.87
제안기법	76.8%	0.991	44.25

실험 결과를 보면, 공격성공률(ASR) 측면에서 제안기법이 가장 좋은 성능을 내었으며, 세부적으로 살펴보면 CPM, VM, C&V(0.2:0.4)와 비교했을 때, 각각 12.5% 포인트, 2.5% 포인트, 2.8% 포인트 향상된 결과를 나타내었다. 지각적 품질 측면에서는 SSIM 기준으로 가장 우수했으며, C&V와 비교시에도 0.03 높았으며, PSNR 측면에서는 0.62 정도 낮았다.

다음으로, 표 2은 공정한 비교를 위해 공격에 활용될 픽셀수를 50%로 제한하여 비교한 결과이다.

[표 2] 비교 실험결과(공정 비교, 50% 픽셀 사용 기준)

기법	ASR	SSIM	PSNR
CPM 단일	56.2%	0.979	47.25
VM 단일	64.5%	0.971	46.47
C&V(0.2:0.4)	65.9%	0.971	46.76
제안기법	76.8%	0.991	44.25

실험 결과를 보면, 공격성공률(ASR) 측면에서 제안기법이 가장 좋은 성능을 내었으며, 세부적으로 살펴보면 CPM, VM, C&V(0.2:0.4)와 비교했을 때, 각각 20.6% 포인트, 12.3% 포인트, 10.9% 포인트 향상된 결과를 나타내었다. 지각적 품질 측면에서는 SSIM 기준으로 가장 우수했으나 [표 1]의 결과보다 그 차이가 감소되었다. 그 이유는 CPM, VM, C&V기법들이 공격대상 픽셀수를 50%로 제한하여 지각적 품질이 향상된 이유이나 그 결과 공격성공률 역시 크게 감소한 결과를 내었다. 반면에, 제안기법은 공격대상 픽셀수를 50%로 제한하였음에도 높은 공격성공률과 높은 지각적 품질을 달성할 수 있었다.

위 결과를 통해 제안기법 3~4단계의 상위 픽셀에 공격 기법을 적용하는 것이 지각적 품질은 유지하면서 공격 성공률을 높일 수 있는 것을 증명하는 연구 효과를 보였다.

5. 결론 및 향후 연구

본 연구를 통해 각각의 픽셀을 제안한 기법을 통해 점수화를 하고 순위를 매겨 높은 순위의 픽셀에 공격기법을 적용하는 방법을 제안하였으며 초도 실험을 통해 실효성을 검증하과 기존 기법과 비교하여 더 우수한 성능을 보임을 확인하였다.

향후에는 문헌연구를 통해 조사된 다수의 지각 불가능 영역 탐색 기법들을 통합하고 제안기법에 따라 이미지 내 지각 불가능 영역을 확장한 후 픽셀의 점수화를 통해 공격기법을 적용하여 더 높은 성공률을 달성할 수 있는 연구를 추가로 수행하여 완성할 예정이다.

참고문헌

[1] Szegedy, Wojcuech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2014.

[2] Jin Li, Ziqiang He, Anwei Luo, Jian-Fang Hu, Z. Jane Wang, Xiangui Kang. “AdvAD: Exploring Non-Parametric Diffusion for Imperceptible Adversarial Attacks,” arXiv:2503.09124v1, 2025

[3] Zhibo Wang, Mengkai Song, Siyan Zheng, Zhifei Zhang, Yang Song, Qian Wang. “Invisible Adversarial Attack against Deep Neural Networks: An Adaptive Penalization Approach,” IEEE Transactions on Dependable and Secure Computing (Volume: 18, Issue: 3), 2019

[4] Francesco Croce, Matthias Hein. “Sparse and Imperceivable Adversarial Attacks,” IEEE/CVF International Conference on Computer Vision (ICCV), 2019

[5] 이륜건, 조영호, “복잡도 지각 모델과 분산 맵을 결합한 적대적 예제 생성 기법 연구,” 국방대학교 석사 학위논문, 2024

[6] Ian Goodfellow, Jonathon Shlens, Christian Szegedy. “Explaining and harnessing adversarial examples,” arXiv:1412.6572, 2014.

[7] Ayberk Aydin, Deniz Sen, Berat Tuna Karli, Oguz Hanoglu , Alptekin Temizel. “Imperceptible Adversarial Examples by Spatial Chroma-Shif,” arXiv:2108.02502v2, 2021

[8] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, Linlin Shen. “Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity,” arXiv:2203.05151v4, 2022