# Metric Learning 기반 Adversarial Example 탐지 가능성에 대한 연구

최석환, 신진명, 김정구, 최윤호*
부산대학교, *부산대학교
daniailsh@pusan.ac.kr, sinryang@pusan.ac.kr, kimjg@pusan.ac.kr, *yhchoi@pusan.ac.kr

# A Study on possibility of detection for adversarial examples based on metric learning

Seok-Hwan Choi, Jinmyeong Shin, Jeong Goo Kim, Yoon-Ho Choi*
Pusan National Univ., *Pusan National Univ.

요 약

딥러닝 모델은 다양한 분야에서 안정적인 성능을 보이지만 입력 이미지에 특정 노이즈를 추가하여 딥러닝 모델의 분류 정확도 감소를 유발하는 Adversarial Example에 매우 취약하다. 이러한 Adversarial Example을 방어하기 위한 기존 연구는 세 범주로 분류할 수 있다. (1) 모델 재학습 기반 방법; (2) 입력 변환 기반 방법; (3) Adversarial Example 탐지 방법. 하지만, Adversarial Example 생성 기법의 발전과 함께 발전하는 모델 재학습 기반 및 입력 변환 기반 방법과 달리 Adversarial Example 탐지 방법은 여전히 이진 분류 방법에 머물러 있다. 본 논문에서는 Metric Learning을 기반으로 한 다중 클래스 Adversarial Example 탐지 기법을 제안한다.

## Ⅰ. 서 론

딥러닝 모델은 다양한 분야에서 활용되고 있으며, 자율 주행 자동차 및 맬웨어 분류와 같은 보안에 민감한 영역에서도 활용되고 있다. 하지만, 딥러닝 모델 활용의 증가와 함께 많은 보안 이슈도 등장하고 있으며, 그 중에서 입력에 사람이 인식 할 수 없는 노이즈를 추가하는 Adversarial Example은 딥러닝 모델의 분류 정확도 감소와 같은 심각한 문제를 야기할 수 있다[1]. 이러한 Adversarial Example을 방어하기 위해 많은 방어 방법이 제안되었으며 주로 세 가지 범주로 분류할 수 있다. (1) 모델 재학습 기반 방법[2]; (2) 입력 변환 기반 방법[3]; (3) Adversarial Example 탐지 방법[4]. 모델 재학습 기반 방법은 딥러닝 모델을 재학습하거나 새로운 모델로 학습하여 Adversarial Example을 방어할 수 있으며, 입력 변환 기반 방법은 Adversarial Example을 딥러닝 모델에 공급하기 전에 노이즈 제거 기법을 적용하여 Adversarial Example을 방어할 수 있다. 이러한 두 가지 방법은 Adversarial Example 생성 기법의 발전과 함께 발전하였다. 반면에, Adversarial Example 자체를 탐지하는 Adversarial Example 탐지 방법은 여전히 이진 분류 방법에 머물러 있다.

따라서 본 논문에서는 Adversarial Example 탐지 방법의 발전을 위해 Metric Learning을 기반으로 한 다중 클래스 Adversarial Example 탐지 기법을 제안한다.

## Ⅱ. 본론

본 논문에서는 다중 클래스 Adversarial Example 탐지를 위해 Convolution Neural Network (CNN) 기반의 유클리드 임베딩 기술을 적용하였으며, 이를 그림 1에서 도시화 하였다. 구체적으로, 제안하는 방법은 정상 입력 및 Adversarial Example을 유클리드 공간에 맵핑하기 위해 학습 데이터셋 생성, Metric Learning 모델 학습, Advesrarial Example 탐지의 3 단계를 거쳐 동작한다.

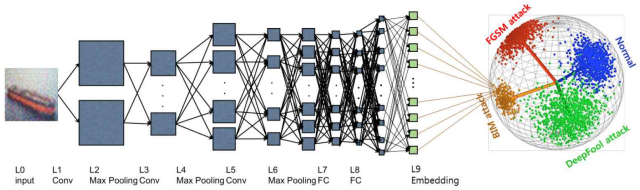



그림 1 제안하는 기법 개요도

### 2.1 학습 데이터셋 생성

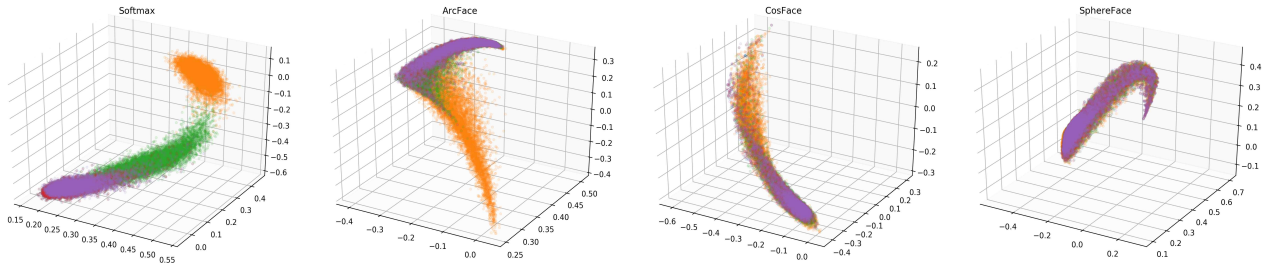
학습 데이터셋 생성 단계에서는 대상 모델에 대해 다양한 Adversarial Example 생성 기법을 적용하여 Metric Learning 모델 학습에 필요한 Adversarial Example을 생성한다. 본 논문에서는 대표적인 Adversarial Example 생성 방법인 Fast Gradient Sign Method (FGSM)[5], Basic Iterative Method (BIM)[6], DeepFool[7], C&W's Method[8]을 이용하여 Metric Learning을 위한 학습 데이터셋을 생성하였다.

### 2.2 Metric Learning 모델 학습

Metric Learning 모델 학습 단계에서는 원본 학습 데이터셋과 Adversarial Example 데이터셋을 사용하여 CNN 기반 Metric Learning 모델을 학습하였다. 본 논문에서는 대표적인 CNN 모델인 ResNet20을 사용하였다. 또한 본 논문에서는 Metric Learning 모델에서 주로 사용되는 Softmax, SphereFace[9], CosFace[10], ArcFace[11]의 손실함수를 이용하여 4개의 Metric Learning 모델을 학습하였다.

### 2.3 Adversarial Example 탐지

Adversarial Example 탐지 단계에서는 학습된 Metric Learning 모델을 이용하여 Adversarial Example을 탐지한다. 학습된 Metric Learning 모델은 입력 이미지를 유클리드 공간 상에 맵핑하므로 학습 데이터셋에




a Softmax 손실 함수 b ArcFace 손실 함수 c CosFace 손실 함수 d SphereFace 손실 함수

그림 2 다양한 손실 함수를 사용한 제안하는 기법의 유클리드 공간 상 맵핑 결과

포함된 Adversarial Example 뿐만 아니라 새로운 유형의 Adversarial Example도 탐지 할 수 있다. Metric Learning 모델 이 후 단계에서는 일반적인 분류 및 클러스터링 알고리즘을 적용할 수 있다. 따라서, 본 논문에서는 Metric Learning 모델을 통한 유클리드 공간 상에 맵핑된 결과만을 다룬다.

### 2.4 실험 및 검증

제안하는 방법의 성능을 평가하기 위해 본 논문에서는 CIFAR-10 이미지 분류 데이터셋에 대해 다양한 Metric Learning 손실 함수를 이용하여 실험하였다. 구체적으로, CIFAR-10 데이터셋의 전체 학습 데이터셋을 이용하여 대상 모델을 학습하였으며, 전체 테스트 데이터셋을 이용하여 Adversarial Example 생성 및 Metric Learning 모델을 학습하였다. 또한 대상 모델 및 Metric Learning 모델로는 ResNet20을 사용하였으며 Metric Learning의 출력 차원은 10차원으로 고정하였다.

그림 3은 다양한 손실 함수를 사용한 제안하는 방법의 결과를 보여준다. Softmax 손실 함수를 사용하여 Metric Learning 모델을 학습한 경우 정상 입력과 4개의 Adversarial Example 생성 기법을 유클리드 공간 상에 효율적으로 맵핑하는 것을 확인할 수 있다. 하지만, SphereFace, CosFace, ArcFace 손실 함수를 사용하여 Metric Learning 모델을 학습한 경우에는 정상 입력 및 4개의 Adversarial Example 생성 기법을 효율적으로 맵핑하지 못하였다. 이는 SphereFace, CosFace, ArcFace 손실 함수가 학습 시에 Adversarial Example 생성 기법의 특징을 반영하지 못한다는 것을 나타낸다.

## Ⅲ. 결론

본 논문에서는 Metric Learning을 기반으로 한 다중 클래스 Adversarial Example 탐지 방법을 제안하였다. 다양한 손실함수를 이용한 실험을 통해 Adversarial Example의 유형을 분류 할 수 있는 가능성을 확인하였다. 하지만, 제안하는 방법은 DeepFool 및 C&W's Method과 같은 유사한 속성을 갖는 Adversarial Example에 대해 낮은 분류 성능을 보였다. 따라서, 향후 연구에서는 데이터 전처리 및 차원 감소 등의 방법을 적용하여 대부분의 Adversarial Example을 효율적인 분류할 수 있는 방안에 대한 연구가 수행되어야 할 것이다.

## ACKNOWLEDGMENT

본 연구는 한국연구재단 논문연구과제 (NRF-2018R1D1A3B07043392) 지원 및 BK21플러스, IT기반 융합산업 창의인력양성사업단의 연구결과로 수행되었습니다

## 참 고 문 헌

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al. “Intriguing properties of neural networks”. In: CoRR abs/1312.6199 (2013). arXiv: 1312 . 6199. URL: http://arxiv.org/abs/1312.6199.

[2] Ruitong Huang, Bing Xu, Dale Schuurmans, et al. “Learning with a Strong Adversary”. In: CoRR abs/1511.03034 (2015). arXiv: 1511.03034. URL: http: //arxiv.org/abs/1511.03034.

[3] Dongyu Meng and Hao Chen. “MagNet: a Two-Pronged Defense against Adversarial Examples”. In: CoRR abs/1705.09064 (2017). arXiv: 1705.09064. URL: http://arxiv.org/abs/1705.09064.

[4] Jiajun Lu, Theerasit Issaranon, and David Forsyth. “SafetyNet: Detecting and Rejecting Adversarial Examples Robustly”. In: The IEEE International Conference on Computer Vision (ICCV). Oct. 2017.

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: International Conference on Learning Representations. 2015. URL: http://arxiv.org/abs/1412.6572.

[6] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio.“Adversarial examples in the physical world”. In: CoRR abs/1607.02533 (2016). arXiv: 1607.02533. URL:http://arxiv.org/abs/1607.02533.

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. “DeepFool: a simple and accurate method to fool deep neural networks”. In: CoRR abs/1511.04599 (2015). arXiv: 1511.04599. URL: http: //arxiv.org/abs/1511.04599.

[8] Nicholas Carlini and David A. Wagner. “Towards Evaluating the Robustness of Neural Networks”. In: CoRR abs/1608.04644 (2016). arXiv: 1608.04644. URL: http://arxiv.org/abs/1608.04644.

[9] Weiyang Liu, Yandong Wen, Zhiding Yu, et al. “Sphereface: Deep hypersphere embedding for face recognition”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 212 - 220.

[10] HaoWang, YitongWang, Zheng Zhou, et al. “Cosface:Large margin cosine loss for deep face recognition”.In: Proceedings of the IEEE Conference on Computer Visionand Pattern Recognition. 2018, pp. 5265 - 5274.

[11] Jiankang Deng, Jia Guo, Niannan Xue, et al. “Arcface:Additive angular margin loss for deep face recognition”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019, pp. 4690 - 4699.