이미지의 강인하지 않은 특징을 이용한 간단한 적대적 기계 학습 방법

임현택, 정기석*
한양대학교

lim3944@hanyang.ac.kr, kchung@hanyang.ac.kr

A Simple Adversarial Training Method Utilizing Non-Robust Feature of Image

Hyun-Tak Lim, Ki-Seok Chung*
Hanyang University, Seoul, Korea

요 약

최근 깊은 신경망(Deep Neural Network), 컨볼루션 신경망(Convolutional Neural Network) 등 인공 신경망을 활용한 인공지능에 대한 연구가 활발히 이루어지고 있다. 특히, 컨볼루션 신경망을 통한 이미지 처리 문제에 대한 발전이 이루어지며 네트워크의 오작동을 유발하는 적대적 공격(Adversarial Attack) 또한 관심을 받고 있다. 이전까지의 연구에서 적대적 공격은 인공 신경망의 선형성과 관련한 문제로 여겨졌는데 최근 연구에서는 이미지에 강인한 특징(Robust Feature)과 강인하지 않은 특징(Non-robust Feature)이 있다는 사실과 강인하지 않은 특징이 적대적 공격에 영향을 미친다는 것이 밝혀졌다. 이에 착안해 본 논문에서는 이미지에서 특징값을 추출하기 위한 가중치 중 강인하지 않은 특징 추출에 관여하는 가중치를 제거해 인공 신경망을 적대적 공격에 강인하도록 재구성하는 방법을 제안했고, 실험을 통해 제안한 방법을 적용한 인공 신경망이 일반 예제에 대해 90.09%의 정확도를, 적대적 공격 예제에 대해 89.13%의 정확도를 보이며 적대적 공격에 대해 강인해지는 결과를 확인할 수 있었다.

Ⅰ. 서 론

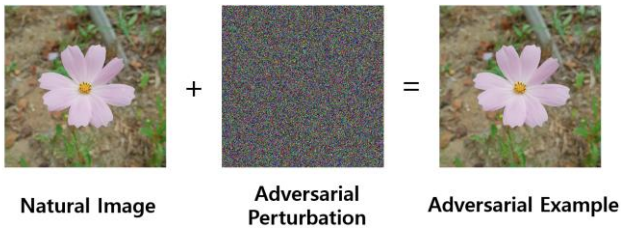
컨볼루션 신경망은 이미지 분류, 물체 인식 등 이미지 처리와 관련한 문제에 대해 뛰어난 성능을 보이며 최근 인공 신경망 연구의 주류를 이루고 있다. 컨볼루션 신경망에 대한 연구가 진행되며 약간의 잡음이 더해진 이미지를 입력으로 해 인공 신경망의 오작동을 유발하는 적대적 공격 또한 대두되었는데, 인공 신경망의 오작동은 자율주행과 같은 실제 사례의 적용에 있어 큰 문제를 불러일으킬 수 있기 때문에 적대적 공격에 강인한 인공 신경망의 필요성이 제시되었다.

이에 따라 적대적 공격에 강인한 인공 신경망을 만들기 위해 적대적 예제를 학습에 이용하는 다양한 학습 기법이 제안되었다. 최근 연구에서는 적대적 공격을 강인한 특징과 강인하지 않은 특징의 관점으로 설명했다[1]. 이에 착안해 본 논문에서는 간단한 방법을 통해 기존의 인공 신경망에서 강인하지 않은 특징을 추출하는 가중치를 제거해 적대적 공격에 강인한 네트워크를 만드는 방법을 제안한다.

Ⅱ. 본론

2.1 적대적 공격

적대적 공격은 인간의 눈으로는 구분할 수 없는 노이즈를 이미지에 섞어 모델의 오작동을 유발하는 인공 신경망 공격 중 하나이다. 대표적인 적대적 공격으로는 Fast Gradient Sign Method(FGSM)이 있다[2]. FGSM 기법은 다음의 수식 1 로 정리된다.



(그림 1) 적대적 공격 예시

$$\eta = \epsilon sign\left(\nabla_x J(\theta, x, y)\right)$$

(수식 1) Fast Gradient Sign Method

수식 1 에서 θ 는 모델의 파라미터, x 는 입력 이미지, y 는 x 의 정답 라벨이고, $J(\theta, x, y)$ 는 인공 신경망 학습을 위한 손실값을 의미한다. FGSM 은 이 cost 의 부호값을 역전파를 통해 입력 이미지에 더해주는 것으로 adversarial example 을 만들어낸다.

이러한 적대적 공격은 MNIST 데이터 셋에 대해 1.6%의 오류율을 보이는 모델에서 99%의 오류율이 나타날 정도로 모델의 심각한 오작동을 유발할 수 있다[2].

2.2 강인한 특징과 강인하지 않은 특징

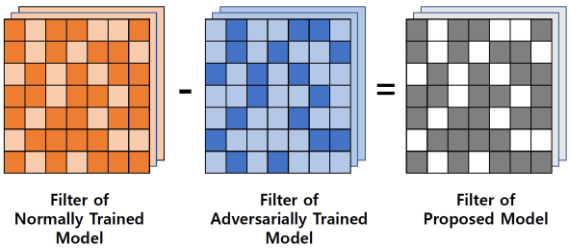
이전까지의 연구에서는 적대적 공격은 인공 신경망의 선형성과 관련되어 발생하는 문제라고 여겨졌지만 최근 연구에서는 인공 신경망이 추출하는 이미지의 특징이 강인한 특징과 강인하지 않은 특징으로 나뉘어 있고, 이 중 강인하지 않은 특징이 적대적 공격에 관여함이

밝혀졌다[1]. 강인한 특징이란 사람이 인식할 수 있는 이미지의 특징으로 고양이 이미지를 예로 들면 귀의 모양, 털의 색상, 꼬리의 길이 등을 강인한 특징이라 할 수 있다. 반대로 강인하지 않은 특징은 사람은 인식하지 못하지만, 인공 신경망은 인식할 수 있는 특징이다.

인공 신경망의 관점에서 강인한 특징은 한 이미지가 자연 상태로 입력됐을 때나 적대적 공격이 포함되어 입력됐을 때나 특징값이 비슷한 수치를 보이는 특징을 의미한다. 반대로 강인하지 않은 특징은 적대적 공격이 포함되어 입력됐을 때 특징값의 수치가 급격히 하강하는 모습을 보인다.

적대적 공격은 정상적인 이미지처럼 보이지만 네트워크의 오작동을 유발하기 때문에 적대적 기계 학습은 이미지의 강인하지 않은 특징을 담당하는 부분에 대해 학습시키는 것으로 생각할 수 있다. 또한, 강인하지 않은 특징에 대한 특징값을 전혀 추출하지 않는다면 모델은 사람이 보는 것과 같이 강인한 특징에 대해서만 추론을 하기 때문에 적대적 공격에 강인해지게 된다. 이에 착안해 본 논문에서는 적대적 기계 학습을 통해 강인하지 않은 특징 부분이 학습된 강인한 네트워크의 가중치를 이용, 일반 네트워크를 강인한 네트워크로 재구성하는 방법을 제안한다.

2.3 제안 방법



(그림 2) 제안 방법 설명

그림 2 에서 색의 진하기로 표현한 것과 같이 일반 학습을 진행한 모델의 경우 강인하지 않은 특징에 관여하는 가중치가 적대적 학습을 진행한 모델에 비해 약하다고 볼 수 있다. 반면, 적대적 학습을 진행한 모델에서는 일반 예제에 대한 정확도가 떨어지기 때문에 일반 학습한 모델에 비해 강인한 특징에 대한 가중치가 약하다. 이러한 점을 활용해 일반 학습을 진행한 모델의 가중치에서 적대적 학습을 진행한 모델의 가중치를 감해주게 되면 일반 학습 모델에서는 강인한 특징에 대한 가중치만 남기 때문에 적대적 공격에 강인해지게 된다.

2.4 실험 결과 및 분석

본 논문에서는 Pytorch 프레임워크를 이용하여 Lenet-5[4]를 실험 모델로, 학습 데이터로는 MNIST 를 이용해 실험을 진행했다. 실험은 먼저 인공 신경망을 MNIST 데이터셋으로 학습을 시킨 네트워크 A 와 MNIST 데이터에 $\epsilon = 0.3$ 의 FGSM 기법을 더해 학습시킨 네트워크 B 를 준비한 후 각각의 일반 데이터 정확도와 적대적 공격에 대한 정확도를 측정했다. 그 후 B 네트워크를 이용해 A 네트워크에서 강인하지 않은 특징의 가중치 부분을 제거해 성능을 측정하였다.

	일반 학습	제안한 방법
일반 예제	99.02%	90.09%
적대적 공격 예제	10.10%	89.13%

(표 1) 제안 방법 실험 결과 (정확도)

표 1 에서 볼 수 있듯이 일반 학습을 진행한 모델은 일반 데이터에 대해서는 높은 정확도를 보였지만 적대적 공격 데이터에 대해서는 정확도가 10% 수준으로 떨어졌다. 이에 반해 적대적 기계 학습을 한 모델을 이용해 제안된 방법으로 가중치를 조정한 모델은 일반 데이터에 대한 정확도의 감소는 있었지만, 적대적 공격에 대해 일반 학습 모델에 비해 8.9 배가량 높은 정확도를 보였다.

Ⅲ. 결론

본 논문에서는 적대적 기계 학습이 진행된 모델을 이용해 일반 학습을 진행한 모델에서 강인하지 않은 특징 추출에 관여하는 가중치를 제거하는 방법을 제안하였다. 제안한 방법을 통해 이미지의 강인한 특징과 강인하지 않은 특징에 관여하는 가중치가 존재함을 확인했으며, 이를 간단한 방법으로 제거해 일반 학습 모델이 적대적 공격에 강인해지는 것을 확인했다.

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2020-0-01304, 모바일 자가 학습 가능 재귀 뉴럴 네트워크 프로세서 기술 개발).

참 고 문 헌

[1] Andrew Ilyas, Shibani Santurkar, et al. “Adversarial Examples are not Bugs, they are Features”, NIPS, 2019

[2] Ian J. Goodfellow, Jonathon Shlens, et al. “Explaining and Harnessing Adversarial Examples”, ICLR, 2015

[3] Aleksander Madry, Aleksandar Makelov, et al. “Towards Deep Learning Models Resistant to Adversarial Attacks", ICLR, 2018

[4] Y. Lecun, et al. "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.