적대적 공격에 견고한 피처 신뢰도 기반 다운샘플링

김우재 ○, 윤성의
카이스트 전산학부
wkim97@kaist.ac.kr, sungeui@kaist.edu

요 약

심층 신경망 기반 분류기에 쓰이는 다운샘플링은 피처의 공간 차원을 줄이는 동시에 유의미한 정보를 보존하는 효과적인 기술이다. 하지만 이미지에 노이즈를 입혀 분류기가 오분류를 하도록 만드는 적대적 공격은 피처 활성화맵을 손상시키고, 이는 피처 활성화 값에 기반하여 출력값을 추출하는 기존의 다운샘플링 기법이 적대적 공격에 취약하다는 것을 나타낸다. 본 연구에서는 적대적 공격에 견고한 다운샘플링 기법을 만들기 위해 각각 피처 뉴런이 모델 예측에 얼마나 기여하는지를 이용하여 피처의 신뢰도를 계산한다. 이 신뢰도를 기반으로 다운샘플링 트리플렛 손실 함수(L_{DTL})를 계산하여 신뢰도가 높은 정보를 추출하고, 신뢰도가 낮은 정보를 무시하도록 합성곱층 기반 다운샘플링 층을 학습시킨다. 최종적으로, 일반 분류기 학습과 적대적 분류기 학습을 통하여 제안 기법의 성능을 확인하였다.

1. 서론

컴퓨터 비전 분야의 태스크에 쓰이는 심층 신경망은 중간 층에서의 공간 차원(spatial dimension)을 줄이기 위해 다운샘플링(downsampling) 기법을 적용한다. 흔히 쓰이는 다운샘플링 기법으로는 풀링(pooling) 혹은 스트라이드(stride)를 이용한 합성곱층(convolutional layer)이 존재한다. 이러한 기법은 전 합성곱층에서 나온 피처(feature)를 입력으로 받아 활성화(activation) 값에 기반하여 출력값을 추출한다는 공통점을 가지고 있다.

하지만 피처 활성화맵은 심층 신경망 기반 분류기에 적대적 교란(adversarial perturbation)을 입혀 잘못된 분류를 하도록 만드는 기법인 적대적 공격(adversarial attack)[3]에 취약하다[5, 9, 12]. 그림 1에서 볼 수 있듯이 적대적 공격은 분류기의 피처 공간을 손상시키고, 기존 이미지의 피처에 없는 노이즈를 입힌다. 이 노이즈는 신경망의 픽셀 공간부터 마지막 로짓(logit) 층까지 쌓여 분류기가 오분류를 하도록 만든다. 따라서, 피처 활성화값에 의존하는 기존의 다운샘플링 기법은 신뢰도가 낮은 정보, 즉 공격에 의해 손상된 정보를 다음 층에 넘겨줄 수 있고, 이는 모델이 올바른 분류를 하는 데 큰 악영향을 끼칠 수 있다.

본 연구에서는 피처 활성화값이 아닌 모델 예측에 대한 피처의 기여도를 기반으로 신뢰도가 높은 정보를 추출하여 적대적 공격에 견고한 다운샘플링 기법을 제안한다. 피처 뉴런이 분류기가 올바른 예측을 하는 데 기여를 많이 한다면, 그 뉴런은 신뢰도가 높은 뉴런일 것이고, 반대로 분류기가 틀린 예측을 하는 데 기여를 많이 한다면, 그 뉴런은 신뢰도가 낮은 뉴런일 것이다. 이 직관에 기반하여 정답 클래스에 기여도가 높은 뉴런을 추출하고, 공격에 의한 오답 클래스에 기여도가 높은 뉴런을 무시하도록 설계된 다운샘플링 트리플렛 손실 함수(L_{DTL})을 제안한다. 그리고 이 손실 함수를 기반으로 합성곱층으로 이루어진 다운샘플링 층을 학습한다. 제안된 다운샘플링 층으로 기존의 피처 활성화값 기반의 다운샘플링 기법을 교체하고 적대적 공격에 대한 분류기의 견고성을 높인다.

2. 관련 연구

다운샘플링 기법과 적대적 공격에 대한 방어 기법은 독립적으로 연구되어 왔다.

2.1 다운샘플링 기법

다운샘플링은 피처의 유의미한 정보를 최대한 많이 유지하며 피처의 차원을 줄이는 데 목표를 두고 있다. 초기의 심층 신경망은[7, 11] 주어진 커널에서 피처 값이 제일 높은 뉴런을 뽑는 최대풀링(max pooling)과 피처 값의 평균을 계산하는 평균풀링 (average pooling)을 사용하였다. ResNet[4] 등의 이후의 신경망은 합성곱층 (convolutional layer)을 사용해 피처의 차원을 줄였다. 그 외에도 랜덤하게 뽑은 피처를 추출하는 Stochastic Pooling[13]과 보조 신경망을 이용해 피처의 중요도를 계산하는 Local

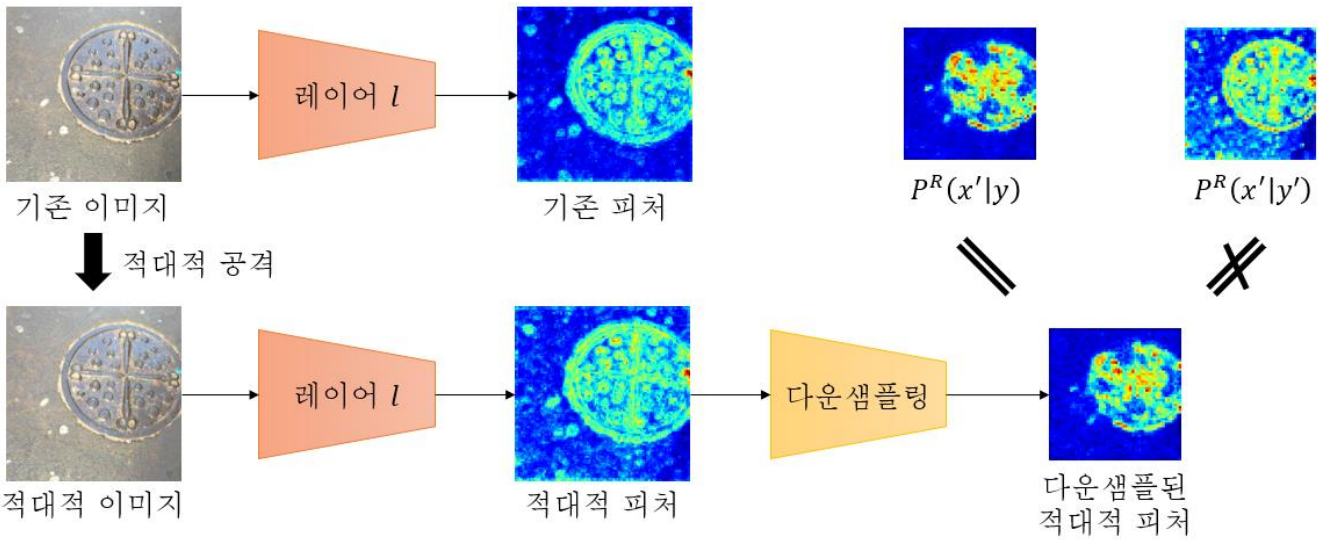

그림 1. 피처 뉴런의 기여도에 기반하여 적대적 공격에 견고한 다운샘플링을 학습하는 기법의 다이어그램. 적대적 공격은 이미지의 피처에 잡음을 입혀 분류기가 오분류를 하게 만든다. 본 연구에선 이러한 적대적 피처에서 신뢰도가 높은 뉴런을 추출한 $P^R(x'|y)$에 흡사하고 신뢰도가 낮은 뉴런을 추출한 $P^R(x'|y')$과 다르도록 학습된 견고한 다운샘플링 기법을 제안한다.

Importance Pooling[2]도 연구가 되었다. 하지만 이러한 기법은 모두 피처의 값에 기반하여 다운샘플링을 하기 때문에, 피처를 손상시키는 적대적 공격에 취약하다.

2.2 적대적 공격에 대한 방어 기법

적대적 공격에 방어하기 위한 가장 효과적인 방법으로는 적대적 이미지(adversarial image)로 분류기를 훈련함으로써 분류기가 적대적 공격에 견고해지도록 만드는 적대적 훈련(adversarial training)이 있다[3, 8]. 하지만 적대적 훈련으로 학습된 분류기도 공격을 받으면 중간 층의 피처가 교란되는 문제점이 있다. 이 문제를 해결하기 위해 피처의 노이즈를 제거하는 기법[12], 피처의 견고함을 정의하는 방법[5], 다른 분류 클래스의 피처 분포를 퍼지게 만드는 연구[9] 등이 진행되어 왔다. 하지만 다운샘플링의 견고함을 살펴본 연구는 진행되지 않았고, 본 논문에서는 적대적 공격에 견고한 다운샘플링 기법을 만들어 적대적 훈련의 성능을 높이려 한다.

3 적대적 공격에 견고한 다운샘플링 기법

본 논문에서는 어떤 피처 뉴런이 신뢰할 수 있는 뉴런인지를 판단하고, 주어진 커널 안에서 신뢰도가 높은 뉴런을 추출하도록 다운샘플링을 제작한다. 뉴런의 신뢰도는 분류기의 예측에 대한 기여도로 나타낸다. 특정 뉴런이 분류기가 올바른 예측을 하는 데 기여도가 높다면 신뢰도 또한 높다고 판단하였고, 반대로 분류기가 틀린 예측을 하는 데 기여도가 높다면 신뢰도 또한 낮다고 판단하였다.

3.1 피처 뉴런의 기여도 계산법

분류기의 예측에 대한 피처 뉴런의 기여도를 계산하기 위해 Layerwise-Relevance Propagation (LRP)[1]를 사용하였다. 층 J 에 있는 임의의 뉴런 a_j의 기여도 R_j는 다음 식으로 계산할 수 있다.

$$R_j = \sum_{k \in K} \frac{a_j w_{jk}}{\sum_{j \in J} a_j w_{jk}} R_k \quad (1)$$

수식의 w_{jk}는 뉴런 a_j와 다음 층 K 에 있는 임의의 뉴런 a_k 사이의 가중치를 의미하고, 분자의 $a_j w_{jk}$는 a_j가 a_k에 가하는 기여도를 의미한다. 분모의 식은 a_j가 속해 있는 층 J 의 모든 뉴런에 대해 같은 방식으로 계산한 기여도의 합이라고 볼 수 있고, 층 J 의 모든 뉴런이 뉴런 a_k에 가하는 기여도의 합이라고 볼 수 있다. 따라서 식의 분수는 층 J 의 모든 뉴런에 비교해 a_j가 a_k에 가하는 기여도의 비율이라고 해석할 수 있고, 이 비율을 뉴런 a_k의 실제 기여도인 R_k로 곱한 값이 a_j가 a_k에 가하는 실제 기여도이다. 이를 층 K 의 모든 뉴런에 가한 합은 뉴런 a_j가 다음 층 K 에 얼마나 많은 기여도를 가하는 지를 나타낸다. R_k의 값은 특정 클래스에 대한 분류기의 제일 마지막 로짓 값으로 설정하고, 역전파를 하며 매 층마다의 기여도를 반복적으로 계산한다.

공격을 받지 않은 이미지가 속하는 정답 클래스를 y, 공격을 받은 이미지가 오분류되는 오답 클래스를 y'이라고 정의했을 때, 신뢰도가 높은 뉴런은 클래스 y 에 높은 기여도를, 그리고 오답 클래스 y' 에 낮은 기여도를 가진다.

3.2 다운샘플링의 구조와 훈련 방법

뉴런의 기여도에 기반하여 신뢰도가 높은 피처

를 다운샘플링 하는 구조와 훈련 방법을 제안한다. 분류기를 테스트할 때는 정답 레이블이 주어지지 않기 때문에 피처에 대한 기여도를 계산할 수 없다. 따라서, 훈련이 가능한 가중치를 지닌 합성곱층으로 다운샘플링 층을 구현하여 신뢰도가 높은 뉴런을 추출하고 신뢰도가 낮은 뉴런을 무시하도록 학습한다. 본 연구에서는 커널 사이즈를 (2, 2), 스트라이드를 2 로 설정하여 겹치는 뉴런 없이 각각 커널에서 피처를 추출하도록 학습한다.

다운샘플링 층을 훈련하기 위해 1) 신뢰도가 높은 정보와 2) 신뢰도가 낮은 정보를 가진 두가지 비교 대상을 정의한다. 두 대상 또한 커널 사이즈 (2, 2)와 스트라이드 2 로 피처를 풀링하여 다운샘플링 층의 산출값과 같은 차원을 지닌다. 임의의 적대적 이미지 x'에 대해서, 정답 클래스 y 에 대한 기여도가 높은 뉴런을 풀링한 값을 신뢰도가 높은 정보로 정의하고 $P^R(x'|y)$로 이름 지었다. 반대로 오답 클래스 y'에 대한 기여도가 높은 뉴런을 풀링한 값을 신뢰도가 낮은 정보로 정의하고 $P^R(x'|y')$으로 이름 지었다. 그림 1 에서도 볼 수 있듯이, $P^R(x'|y)$는 적대적 이미지의 피처를 풀링하였지만 신뢰도가 높은 정보를 추출하였기 때문에 공격을 받지 않은 이미지의 피처와 유사한 것을 확인할 수 있다. 반대로 $P^R(x'|y')$은 신뢰도가 낮은 정보를 추출하였기 때문에 적대적 이미지의 손상된 피처에서 생겨난 노이즈를 담고 있는 것을 확인할 수 있다.

적대적 공격에 견고한 다운샘플링 층을 학습하기 위해 비교 대상 $P^R(x'|y)$와 $P^R(x'|y')$을 이용하여 다운샘플링 트리플렛 손실 함수(L_{DTL})을 제안한다. 손실 함수의 수식은 아래와 같다.

$$L_{DTL} = \max\left(\| P(x') - P^R(x'|y) \|^2 - \| P(x') - P^R(x'|y') \|^2 + \alpha, 0\right) \quad (2)$$

위 수식은 기존의 트리플렛 손실 함수(triplet loss)[10]와 유사한 형태를 지닌다. 적대적 이미지 x'에 대해서 학습하고자 하는 앵커(anchor), 즉 다운샘플링 층의 산출값을 P(x')로 정의하고, 신뢰도가 높은 $P^R(x'|y)$와 신뢰도가 낮은 $P^R(x'|y')$을 각각 긍정(positive) 예제와 부정(negative) 예제로 정의하였다. 그림 1 에서 볼 수 있듯이 앵커와 긍정 예제의 거리를 줄이고, 앵커와 부정 예제의 거리를 늘림으로써 신뢰도가 높은 뉴런을 추출하고 신뢰도가 낮은 뉴런을 무시하도록 학습하였다. 하이퍼마라미터 α 는 긍정 예제와 부정 예제 사이의 여백을 뜻하고, 본 연구에선 0.01 로 설정하였다. L_{DTL} 의 가중치를 하이퍼파라미터 λ 로 조정하고 분류기 학습에서 사용하는 크로스 엔트로피 손실 함수와 혼합하여 훈련을 진행하였다.

4 실험 결과 및 분석

본 연구에서 제안한 적대적 공격에 견고한 다운샘플링 기법의 성능 확인을 위하여 CIFAR-10[6] 데이터셋 대하여 실험 및 분석을 진행하였다. 분류기는 VGG16[11]을 활용하였고, SGD optimizer 를 이용해 100 에폭 동안 학습하였다. 처음에는 0.01 의 러닝 레이트로 시작해 75 에폭과 90 에폭에는 러닝 레이트를 10 씩 나누었다. 비교 모델로는 (1) 최대풀링을 사용한 분류기(**Max-VGG**), (2) 최대풀링을 합성곱층 기반 다운샘플링으로 교체한 분류기(**Conv-VGG**), 그리고 (3) Conv-VGG 의 다운샘플링 층을 본 연구에서 제안한 견고한 다운샘플링 기법, 즉 L_{DTL} 으로 학습한 분류기(**Ours**)를 사용하였다. 성능 비교를 위한 공격으로는 FGSM[3]과 PGD-20, PGD-100[8]을 사용하였고, 적대적 학습을 위한 공격으로는 PGD-10[8]을 사용하였다.

4.1 제안 기법의 견고성 분석

공격을 받지 않은 이미지로 분류기를 학습하여 제안 기법의 견고성을 비교하였다. 표 1 은 공격을 받지 않은 이미지와 다양한 공격을 받은 이미지에 대한 세가지 모델의 분류 정확도를 나타낸다. 적대적 공격에 견고한 다운샘플링 기법을 쓴 Ours 방법이 Max-VGG 와 Conv-VGG 와 비교해서 모든 공격에 가장 견고한 모습을 보였다. 특히 FGSM 공격에서는 Conv-VGG 에 비교해서 분류기의 정확도가 4.69%p 만큼 향상된 것을 확인할 수 있었다. 공격에 대한 견고성 향상을 통해 본 연구에서 제안한 다운샘플링 기법의 효과를 입증할 수 있었다.

표 1. 기존 분류기 학습에서 제안 기법의 성능. 빨간색 글씨는 Conv-VGG 와 비교해서 Ours 의 성능 향상을 나타낸다.

	No attack	FGSM	PGD-20	PGD-100
Max-VGG	90.60%	14.78%	3.64%	4.38%
Conv-VGG	91.40%	15.91%	5.58%	6.82%
Ours	**91.51%**	**20.60%** +4.69%p	**8.85%** +3.27%p	**9.50%** +2.68%p

다음은 공격을 받은 이미지로 분류기를 적대적 학습으로 훈련하여 제안 기법의 성능을 비교하였다. 표 2 는 공격을 받지 않은 이미지와 다양한 공격을 받은 이미지에 대한 세가지 적대적 학습 모델의 분류 정확도를 나타낸다. 기존 분류기 학습 때와 마찬가지로 견고한 다운샘플링 기법을 사용한 Ours 방법이 모든 공격에 가장 견고한 모습을 보였다. 예를 들어 PGD-20 공격에서는 Conv-VGG 에 비교해서 적대적 학습 분류기의 정확도가 0.57%p 만큼 향상된 것을 확인할 수 있었다. 적대적 학습을 한 분류기의 경우 기존 분류기보다 Ours 기법의 성능 향상 폭이

크지 않았는데, 이는 적대적 학습을 한 분류기가 그렇지 않은 분류기보다 공격에 대해 더 견고함을 보이고, 따라서 공격을 받았을 때 덜 손상된 피처를 산출하기 때문이라고 추측된다.

표 2. 적대적 분류기 학습에서 제안 기법의 성능. 빨간색 글씨는 Conv-VGG 와 비교해서 Ours 의 성능 향상을 나타낸다.

	No attack	FGSM	PGD-20	PGD-100
Max-VGG	77.51%	50.77%	46.44%	47.96%
Conv-VGG	**78.03%**	51.15%	47.19%	48.61%
Ours	77.25%	**51.50%** +0.35%p	**47.76%** +0.57%p	**49.09%** +0.48%p

4.2 하이퍼마라미터 변화에 따른 성능 비교

그림 2 에서 볼 수 있듯이, 다운샘플링 트리플렛 손실 함수(L_{DTL})의 가중치인 하이퍼파라미터 λ 에 따른 성능 변화를 적대적 학습 분류기의 분류 정확도로 비교하였다. 가중치 λ 가 0.01 일 때, 모든 공격(FGSM, PGD-20, PGD-100)에 대해서 가장 높은 견고함을 보였다. λ 가 0.1 로 더 클 경우에는, 견고함이 오히려 떨어지는 모습을 보였다. 이를 통해 L_{DTL} 이 적대적 학습 분류기가 공격에 더 강인해지는 데 도움을 주지만, 그 가중치가 너무 크면 분류기가 크로스 엔트로피 손실 함수를 학습하는 데 악영향을 끼쳐 오히려 정확도를 떨어뜨리는 것을 확인할 수 있었다.

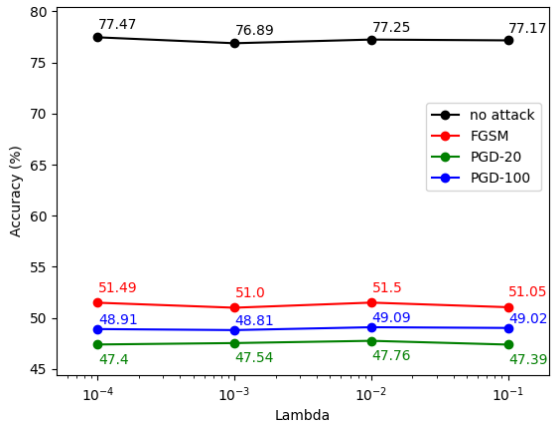

그림 2. 하이퍼파라미터 λ 에 따른 성능 비교.

5 결론

본 연구에서는 피처를 손상시키는 적대적 공격에 견고한 다운샘플링 기법을 제시하였다. 피처 활성화값이 아닌 피처 뉴런의 기여도에 기반해 피처의 신뢰도를 판단하고, 신뢰도가 높은 정보를 추출하도록 다운샘플링을 학습시켰다. 제시한 다운샘플링 기법으로 분류기를 학습시켰을 때, 모든 공격에서 더 높은 견고성을 얻었다.

하지만 제시 방법으로 얻을 수 있는 성능 향상이 한정적이고, 이 문제는 적대적 훈련 시에 더 두드러진다. 이를 해결하기 위해 다운샘플링 기법을 업샘플링(upsampling) 기법과 합쳐 하나의 모듈로 분류기의 더 다양한 피처에서 신뢰도가 높은 정보를 추출하고 견고성을 높일 계획에 있다.

감사의 글

이 성과는 2021 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2019R1A2C3002833).

참고문헌

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PloS one*, 2015.

[2] Ziteng Gao, Limin Wang, and Gangshan Wu. LIP: Local importance-based pooling. In *ICCV*, 2019.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[5] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[9] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *ICCV*, 2019.

[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, 2019.

[13] Matthew D Zeiler and Robert Fergus. Stochastic pooling for regularization of deep convolutional neural networks. In *ICLR*, 2013.