한국정보전자통신기술학회논문지(jkiiect)'24-4, Vol.17 No.2
논문 24-17-02-113 http://dx.doi.org/10.17661/jkiiect.2024.17.2.113

컨벌루션 신경망 모델의 적대적 공격에 따른 성능과 개체군 희소 지표의 상관성에 관한 경험적 연구

이영석*

Empirical Study on Correlation between Performance and PSI According to Adversarial Attacks for Convolutional Neural Networks

Youngseok Lee*

요 약 개체군 희소 지표는 인공 신경망을 구성하고 있는 내부 레이어의 동작을 뉴런의 관점에서 관찰할 수 있기 때문에 블랙박스로 불리는 인공 신경망 내부의 동작을 설명하기 위하여 활용될 수 있다. 최근의 연구에서는 개체군 희소 지표를 두 종류의 컨벌루션 신경망 모델 분석에 적용하여, 레이어의 층이 깊어질수록 지표 값이 비례하여 증가하는 것이 관찰되었음을 보고하였다. 또한, 영상 분류를 위한 컨벌루션 신경망 모델에서 개체군 희소성 지표와 성능이 양의 상관성을 보인다는 연구도 있다. 본 연구에서는 적대적 예제가 컨벌루션 신경망에 적용되었을 때 신경망 내부에서 어떠한 동작이 수행되는지에 대하여 관찰하였다. 이를 위하여 적대적 예제를 입력으로 하는 컨벌루션 신경망의 개체군 희소 지표를 구한 다음, 컨벌루션 신경망의 성능과의 상관성을 비교하였다. 실험의 결과로부터 사전에 5%의 정확도를 갖도록 변형된 적대적 예제들에 대하여 온건한 데이터를 적용한 경우와 유사한 패턴의 양의 상관성을 갖는 것을 확인할 수 있었다. 이 실험결과는 적대적 예제와 온건한 데이터에 대한 각각의 개체군 희소성 지표 값들이 거시적인 관점에서 차이가 없다는 것을 의미하며 적대적 예제가 뉴런의 활성화 측면에서부터 적대적으로 동작한다는 것을 의미한다.

Abstract The population sparseness index(PSI) is being utilized to describe the functioning of internal layers in artificial neural networks from the perspective of neurons, shedding light on the black-box nature of the network's internal operations. There is research indicating a positive correlation between the PSI and performance in each layer of convolutional neural network models for image classification. In this study, we observed the internal operations of a convolutional neural network when adversarial examples were applied. The results of the experiments revealed a similar pattern of positive correlation for adversarial examples, which were modified to maintain 5% accuracy compared to applying benign data. Thus, while there may be differences in each adversarial attack, the observed PSI for adversarial examples demonstrated consistent positive correlations with benign data across layers.

Key Words : Population sparse index, Convolutional neural network, Adversarial attack, Layer

1. 서론

인공 신경망은 인간의 대뇌 피질을 구성하는 뉴런들의 연결로부터 영감을 받아 사람이 외부의 자극에 대하여 인지하고, 판단하며 처리하는 과정을 컴퓨터 연산을 이용하여 모방하는 시스템이다[1]. 시스템의 특성에 따라 인공 신경망은 입력과 출력이 존재하지만, 일반적인 선형 시스템과 달리 비선형적 특성으로 인하여 내부 동작이 블랙박스로 묘사된다. 따라서 인공 신경망 모델의 내부 작동 기전을 밝히는 것은 인공 신경망이

This paper is supported by Chungwoon University Research Fund in 2023
*Corresponding Author : Dept. of Electronic Engineering, Chungwoon University (yslee@chungwoon.ac.kr)
Received January 31, 2024 Revised March 03, 2024 Accepted March 15, 2024

동작하는 원리를 설명할 수 있을 뿐만 아니라 나아가 인간의 대뇌 피질이 어떤 방식으로 외부의 자극에 반응하고 이를 처리하는지에 대한 근본적인 원리를 제시할 수 있는 근거로 작용한다[2,3].

최근의 연구들은 인공 신경망의 내부 동작을 설명하기 위하여 신경과학에서 정의된 지표 및 파라미터를 사용하는 경향이 두드러진다. 이러한 경향 중에 한가지로서, 인공 신경망의 일종인 컨벌루션 신경망의 각 레이어를 구성하는 뉴런들의 활성화 정도를 알 수 있는 척도인 개체군 희소 지표 (PSI: Population Sparseness Index)를 이용하여 분석하는 방법이 최근에 시도되었다[4].

본 연구에서는 컨벌루션 신경망이 적대적 공격을 받는 경우, 각 레이어들의 개체군 희소 지표를 계산하고 성능과 개체군 희소 지표 사이의 상관성을 관찰하였다. 이와 같은 경험적 관찰은 인공 신경망이 적대적 예제에 의하여 오동작할 때에 신경망 내부의 레이어에서 어떤 상황이 벌어지고 있는지 추적할 수 있는 단서를 제공할 수 있다.

2. 관련 연구들

최근 [5]의 연구에 따르면 영상 객체를 인식하기 위한 컨벌루션 신경망 모델의 각 레이어에서 객체의 인식에 관여하는 뉴런들을 표현하는 식 (1)의 개체군 희소성 지표가 신경망의 각 레이어를 거치면서 증가한다는 것을 밝혀내었다. 이 연구는 비록 인공 신경망이 기존의 기계학습 알고리듬들과 비교하여 탁월한 성능을 보이지만, 어떤 과정을 통하여 이와 같은 추론이 가능한지에 대하여 단서를 남겼다는 점에서 의미가 있다.

$$PSI = \frac{1-a}{1-\frac{1}{N_u}} \tag{1}$$

위 식에서 N_u는 각 레이어에 있는 뉴런들의 수이고, a는 각 레이어에 있는 뉴런들의 수와 객체 인식을 위하여 강 범주에 대응하여 활성화되는 뉴런들의 수에 대한 비율[4]로서 뇌가 자극을 인지하고 처리하는 과정에 기여하는 뉴런들에 대한 전체 뉴런들의 참여 비율, 즉 개체군 희소성 지표가 적어도 컨벌루션 신경망 모델에서는 각 레이어에 소속된 뉴런들의 활동을 추정할 수 있는 지표로써 사용할 수 있다는 것을 의미한다. 또한 [5]의 연구에서는 각 컨벌루션 신경망 모델의 레이어에서 추출한 개체군 희소 지표와 해당 레이어의 분류 성능이 특정 레이어 이후로 양의 상관성을 갖는다고 주장하였다.

또한 최근의 연구 [6]에서는 [5]의 연구 결과를 발전시켜 적대적 공격의 일종인 FGSM 공격[7]으로 왜곡된 데이터를 컨벌루션 신경망에 주입하여 얻어지는 개체군 희소 지표를 분석하고, 정상적인 데이터 및 적대적 예제에 의해 얻어지는 개체군 희소 지표의 값이 서로 다른 것을 확인하였다.

본 연구에서는 [5]와 [6]의 연구를 확장하여 적대적 공격들과 데이터셋의 종류들을 확장하여 적대적 예제에 대한 [5]의 주장을 확인하려 한다. 대표적 컨벌루션 신경망 모델인 AlexNet[8]과 VGG11[9]에 대하여 Caltech 256[10] 데이터셋을 적용할 때, 정상적인 데이터셋과 FGSM, PGD[7] 그리고 CW[7] 공격으로부터 얻어진 적대적 예제들을 적용하여 얻어진 개체군 희소 지표와 성능 간의 상관성에 대하여 실험한다. 적대적 예제들은 정상적인 데이터셋에 의하여 학습된 신경망 모델들이 올바른 목표에 접근하는 것을 방해하는 요소로 작용하기 때문에 신경망 모델들의 성능에 영향을 미친다. [5]의 방식과 같은 과정으로 적대적 예제를 적용하는 경우 인공 신경망들은 바르지 않은 결과를 낼 수밖에 없다.

이를 위하여 실험에서는 각각의 적대적 예제를 생성하는 과정에서 적용된 적대적 공격들이 5%의 성능 즉, 두 신경망 모델들이 95%의 확률로 올바르지 않은 결과를 내도록 설정하였다. 따라서 [5]의 실험 과정을 반복하는 경우 적대적 예제에 대하여 [5]의 경우에 유사한 결과를 나타낸다는 것을 예측할 수 있으며 이와 같은 예측은 적대적 예제가 정상적인 데이터와 유사한 개체군 희소 지표를 갖지만, 정상적인 결과와 다른 추론을 한다는 것을 의미한다.

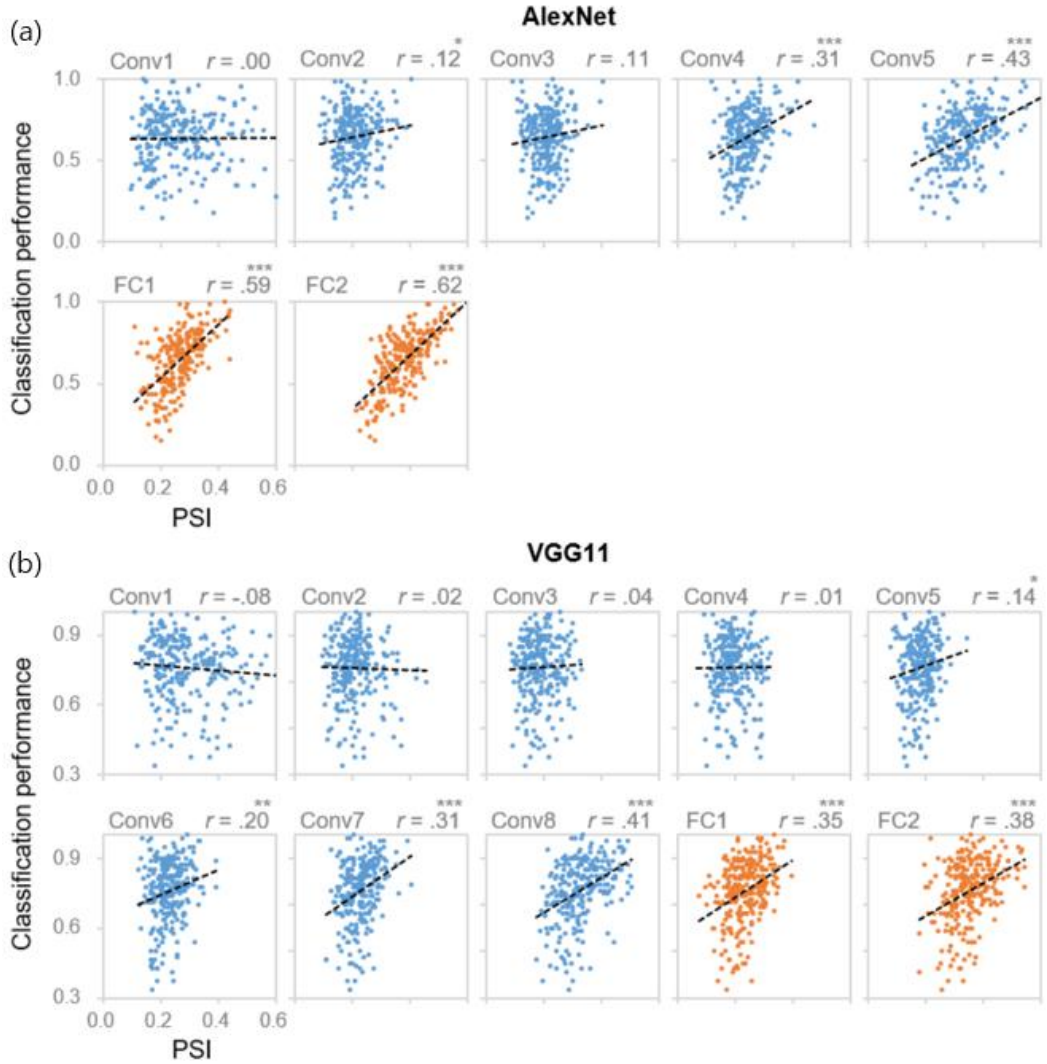

그림 1.컨벌루션 신경망 모델의 레이어들과 개체군 희소 지표간의 상관성 그래프. A. AlexNet, B VGG11

Fig. 1. Correlation graph of PSI vs. performance in convolutional neural networks. (a) AlexNet (b) VGG11

3. 실험 및 결과 고찰

CNN 모델의 각 레이어에 영향을 미치는 뉴런들을 분석하기 위하여 Alexnet[8]와 VGG11[9] 모델을 사용하였다. 두 모델들은 모두 파이토치로 구성된 사전 훈련된 모델을 사용하여 실험을 수행하였다. 그림 1은 Caltech256 데이터셋을 통하여 훈련한 AlexNet과 VGG11 모델에 온건한 실험 데이터를 적용한 결과를 나타내고 있다. 두 모델의 각 레이어에서 개체군 희소 지표와 성능 사이를 스캐터 형식으로 나타내고 1차 함수의 기울기를 이용하여 표현하였다. 두 모델에서 나타나는 공통점은 상관성을 표현하는 1차 근사화 함수의 기울기가 레이어를 거치면서 점점 증가한다는 것이다. 예를 들어 AlexNet의 경우에는 컨벌루션 레이어 4번에서 0.31의 상관계수를 나타내지만 레이어를 거치면서 값이 점점 증가하여 최종

적으로 전 연결 레이어(fully connected layer) 2번에서는 0.62의 상관계수를 갖는 것을 확인할 수 있다. 상관성을 표현할 수 있는 다른 방법으로는 1차 근사화 함수의 기울기를 각도로 표현하는 것이다. 이 경우에도 기울기의 각도가 점진적으로 증가하는 것을 관찰할 수 있다.

그림 2는 FGSM 공격으로 생성된 적대적 예제들을 AlexNet과 VGG11 모델에 적용하고 관찰한 개체군 희소 지표와 성능의 상관성을 그래프를 나타내고 있다. AlexNet 모델의 경우에는 컨벌루션 레이어 5번에서 1차 근사화 함수의 기울기가 이전 레이어 보다 감소한 것이 관찰되어 온건한 데이터를 사용한 그림 1의 결과와 차이점을 보였으며 전 연결 레이어 1번보다 전 연결 레이어 2번에서 상관성을 나타내는 1차 근사화 함수의 기울기가 감소하는 것을 관찰할 수 있다.

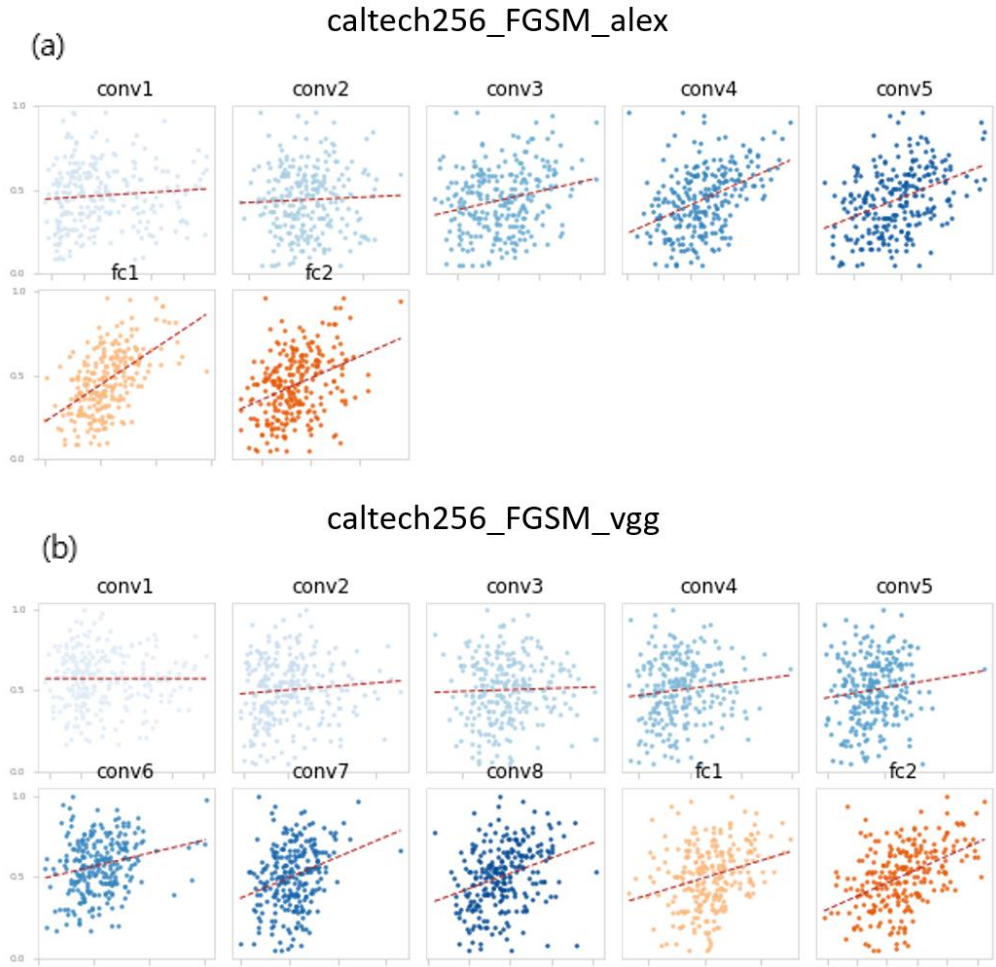

그림 2. FGSM 공격에 따른 컨벌루션 신경망 모델의 레이어들과 개체군 희소 지표간의 상관성 그래프.
(a) AlexNet (b) VGG11

Fig. 2. Correlation graph of PSI vs. performance in FGSM-attacked convolutional neural networks.
(a) AlexNet (b) VGG11

표 1.AlexNet에 대한 온건한 데이터와 공격을 받은 데이터의 성능과 개체군 희소 지표 간의 상관성 기울기 비교
Table 1. Correlation slope comparison of performance vs. PSI in benign and attacked data for AlexNet

Attacks	CALTECH256_AlexNet		IMAGENT_AlexNet	
None	conv1	-1.3	conv1	-7.6
	conv2	9.1	conv2	18.4
	conv3	8.6	conv3	8.3
	conv4	21.8	conv4	13.2
	conv5	23.5	conv5	21.1
	fc1	28.2	fc1	29.1
	fc2	35.4	fc2	37.3
FGSM	conv1	8.5	conv1	3.5
	conv2	4.4	conv2	8.3
	conv3	13.2	conv3	5.2
	conv4	23.3	conv4	10.1
	conv5	21.5	conv5	8.5
	fc1	32.9	fc1	0.1
	fc2	23.8	fc2	-5.0
PGD	conv1	7.5	conv1	-3.5
	conv2	3.3	conv2	11.9
	conv3	12.2	conv3	17.8
	conv4	27.2	conv4	20.7
	conv5	23.3	conv5	30.1
	fc1	33.9	fc1	30.3
	fc2	27.8	fc2	35.2
CW	conv1	8.6	conv1	-5.5
	conv2	-2.7	conv2	18.6
	conv3	13.0	conv3	10.9
	conv4	8.1	conv4	17.8
	conv5	18.3	conv5	25.2
	fc1	12.4	fc1	28.5
	fc2	-10.5	fc2	40.4

표 2. VGG11에 대한 온건한 데이터와 공격을 받은 데이터의 성능과 개체군 희소 지표 간의 상관성 기울기 비교
Table 2. Correlation slope comparison of performance vs. PSI in benign and attacked data for VGG11

Attacks	CALTECH256_VGG11		IMAGENT_VGG11	
TYPICAL	conv1	-9.8	conv1	-5.1
	conv2	-3.2	conv2	-2.5
	conv3	3.3	conv3	11.0
	conv4	1.6	conv4	11.2
	conv5	11.5	conv5	11.5
	conv6	8.5	conv6	11.9
	conv7	21.9	conv7	15.8
	conv8	20.8	conv8	18.4
	fc1	20.5	fc1	25.5
	fc2	20.7	fc2	23.8
FGSM	conv1	0.1	conv1	2.5
	conv2	5.3	conv2	3.6
	conv3	3.5	conv3	7.2
	conv4	9.9	conv4	11.2
	conv5	11.7	conv5	13.5
	conv6	13.6	conv6	20.3
	conv7	22.2	conv7	15.6
	conv8	20.4	conv8	22.9
	fc1	18.2	fc1	29.0
	fc2	25.3	fc2	33.4
PGD	conv1	2.5	conv1	-3.2
	conv2	7.4	conv2	1.5
	conv3	3.6	conv3	8.7
	conv4	9.5	conv4	10.6
	conv5	12.4	conv5	18.3
	conv6	13.3	conv6	17.9
	conv7	22.9	conv7	19.8
	conv8	19.8	conv8	20.7
	fc1	13.5	fc1	30.7
	fc2	25.7	fc2	35.0
CW	conv1	7.5	conv1	-3.1
	conv2	7.9	conv2	-1.5
	conv3	3.6	conv3	10.5
	conv4	9.8	conv4	13.9
	conv5	12.7	conv5	11.7
	conv6	8.6	conv6	16.6
	conv7	7.9	conv7	18.4
	conv8	19.7	conv8	13.3
	fc1	12.6	fc1	24.5
	fc2	15.4	fc2	28.9

또한 VGG11 모델에서는 컨벌루션 레이어 8번의 기울기가 전 레이어에 비하여 감소하는 것이 관찰되었으나, 감소하는 정도는 온건한 데이터보다 FGSM 공격을 받은 데이터에서 더 뚜렷하게 관찰되었다.

FGSM 공격과 동일한 실험이 PGD 및 CW 공격으로부터 생성된 적대적 예제에 대해서도 수행되었다. 두 적대적 공격에 대한 실험은 FGSM 적대적 예제와 동일한 실험환경, 즉 공격 강도로 추론 정확도를 조정할 수 있는 적대적 예제 생성 소프트웨어[11]를 이용하여 수행되었으며, 본 연구에서는 모든 적대적 공격들의 추론 정확도를 5% 이하가 되도록 실험을 설정하여 AlexNet과 VGG11 모델이 부정확하게 추론을 유도하도록 하는 적대적 예제를 생성하였다. 또한, 온건한 데이터는 Caltech 256 및 ImageNet 데이터셋에서 학습을 위한 데이터들을 제외한 나머지 데이터들로부터 수집하였고, 이 데이터들은 AlexNet과 VGG11 모델의 타당도 평가에서 90% 이상의 추론 정확도를 나타내었다.

표 1은 Catech 256 데이터셋에 대하여 온건한 데이터와 적대적 예제들을 AlexNet 모델에 적용한 결과를 보여 주고 있다. 표 1에서 None으로 표시된 부분은 공격을 받지 않은 온건한 데이터에 대한 실험결과를 나타내고 있다. FGSM 공격의 경우 온건한 데이터에 해당하는 개체군 희소 지표와 성능 사이의 상관성은 레이어가 진행됨에 따라 점진적으로 증가하는 반면, FGSM 공격을 받은 데이터의 상관성은 AlexNet의 경우에는 점진적으로 증가하다가 컨벌루션 레이어 8번 이후로 일정한 수준을 유지하는 경향을 나타내었으며 VGG11의 경우에는 컨벌루션 레이어 7번에서 가장 큰 상관성을 보이고 이후 전 연결 레이어 1까지 감소하다가 마지막 레이어인 전 연결 레이어에서 증가하는 것이 관찰되었다. 나머지 두 공격에 대해서도 레이어가 증가할수록 개체군 희소 지표의 점진적 증가가 관찰되었으며 Caltech 256 데이터셋의 개체군 희소 지표 값이 ImageNet 데이터셋의 개체군 희소 지표 값보다 더 큰 값을 나타내었다.

표 2는 VGG11 모델에 대하여 동일한 실험을 수행한 결과를 나타낸다. Caltech 256 데이터셋과 ImageNet 데이터셋에 대한 개체군 희소 지표가 레이어가 증가함에 따라 점진적으로 증가하는 것을 관찰할 수 있으며 ImageNet 데이터셋의 결과가 Caltech256 데이터셋의 결과보다 개체군 희소성 지표의 값이 큰 것을 관찰할 수 있다.

즉, 표 1과 표2의 결과로부터 두가지 훈련 데이터셋으로 훈련한 컨벌루션 신경망 모델, AlexNet과 VGG11에서 온건한 데이터와 적대적 예제에 대한 개체군 희소 지표와 성능 사이의 상관성은 거시적인 관점에서 양의 상관성을 보이며 레이어의 진행에 따라 점진적으로 증가하는 패턴을 나타내었고, 미시적인 관점에서는 상관성을 나타내는 1차 근사화 함수의 기울기 각도에서 미소한 차이를 나타내는 것이 관찰되었다.

두 컨벌루션 모델에서 온건한 데이터와 적대적 예제에 대한 상관성이 거시적으로 양의 상관성이 증가하는 추세를 갖는 가운데 주목할 만한 두 데이터 사이에 차이점이 각각의 모델에 존재한다. 첫 번째 차이점은 온건한 데이터가 사용된 AlexNet 모델에서는 컨벌루션 레이어 3번 이후로 상관성이 점진적으로 증가하는 패턴을 보이지만, 적대적 예제가 사용된 AlexNet 모델은 컨벌루션 레이어 5번과 전 연결 레이어 2번에서 이전 레이어의 기울기의 각도에 비하여 상관성이 감소하는 것이 관찰되었다.

특히, 마지막 전 연결 레이어는 이전 레이어에 비하여 상관성이 감소하는 것이 관찰되었다. VGG11 모델에서는 온건한 데이터가 모델에 적용된 경우 컨벌루션 레이어 5번 이후로 기울기 각도가 점진적으로 증가하거나 값을 유지하는 경향이 있지만, 적대적 예제가 적용된 경우에는 컨벌루션 레이어 7번부터 지속적인 값의 감소 이후, 마지막 전 연결 층에서 값이 증가하는 패턴을 나타내었다.

이와 같은 결과는 컨벌루션 신경망 모델의 내부에서 적대적 예제에 의하여 정상적인 추론 과정과 다른 패턴의 추론이 이루어지고 있음을 알 수 있다. 즉, 신경망 모델에 적대적 예제가 적용되었을 때 개체군 희소 지표와 성능의 상관성을 나타내는 값을 이용하여 적대적 예제의 존재를 추정할 수 있는 유의미한 근거를 제공할 수 있다.

4. 결론

본 연구에서는 신경생리학 분야에서 뉴런의 활동을 관찰하기 위하여 사용한 개체군 희소 지표 분석을 통하여 대표적인 두 종류의 컨벌루션 신경망 모델인 Alexnet과 VGG11에 대한 각 레이어별 성능과 뉴런의 활동에 대하여 관찰하는 실험 및 분석을 수행하였다. 두 모델에 대하여 FGSM, PGD 및 CW 공격으로부터 얻어진 5% 이하의 추론 정확도를 갖는 적대적 예제와 온건한 데이터들로 이루어진 정상적인 데이터셋 사이에 레이어별 개체군 희소 지표 분석에서 두 데이터셋 모두 레이어가 증가함에 따라 성능과 개체군 희소 지표가 양의 상관성을 갖는 것을 관찰할 수 있다. 또한, 두 데이터셋에서 적대적 예제의 개체군 희소 지표의 값이 온건한 데이터에 대한 개체군 희소성 지표의 값에 비하여 더 큰 것을 관찰 할 수 있는데, 이와 같은 결과는 CNN 모델의 각 레이어에서 사용하는 뉴런의 개수가 적대적 예제에서 상대적으로 적은 것을 의미한다. 이와 같은 결과는 신경망 모델에 입력되는 데이터의 적대적 여부를 판단할 수 있는 근거 자료로 이용할 수 있을 뿐만 아니라, 시각 피질 연구와 관련하여 착시(optical illusion)를 인공 신경망의 관점에서 해석할 수 있는 가능성을 제시한다.

REFERENCES

[1] Beniaguev, David, Idan Segev, and Michael London. "Single cortical neurons as deep artificial neural networks." Neuron 109.17, 2727-2739, 2021.

[2] Lau, Brian, Garrett B. Stanley, and Yang Dan. "Computational subunits of visual cortical neurons revealed by artificial neural networks." Proceedings of the National Academy of Sciences 99.13, 8974-8979, 2002.

[3] Shi, Jianghong, Eric Shea-Brown, and Michael Buice. "Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex." Advances in Neural Information Processing Systems 32, 2019.

[4] Lehky, Sidney R., Terrence J. Sejnowski, and Robert Desimone. "Selectivity and sparseness in the responses of striate complex cells." Vision research 45.1, 57-73, 2005.

[5] Liu, Xingyu, Zonglei Zhen, and Jia Liu. "Hierarchical sparse coding of objects in deep convolutional neural networks." Frontiers in computational neuroscience 14, 578158.

[6] Youngseok Lee, "Study on the Neural Activities for Adversarial Examples in Convolutional Neural Network model by Population Sparseness Index", J. of KIIECT, 16-1, 1-7, 2023.

[7] Goodfellow Ian, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples", arXiv preprint arXiv:1412.6572, 2014.

[8] Yuan, Zheng-Wu, and Jun Zhang. "Feature extraction and image retrieval based on AlexNet." Eighth International Conference on Digital Image Processing (ICDIP 2016). Vol. 10033. SPIE, 2016.

[9] Yu, Wei, et al. "Visualizing and comparing AlexNet and VGG using deconvolutional layers." Proceedings of the 33rd International Conference on Machine Learning. 2016.

[10] Griffin, Gregory, Alex Holub, and Pietro Perona. "Caltech-256 object category dataset." 2007.

[11] http://adversarial-robustness-toolbox.readthedocs.io

저자약력

이 영 석 (Young-Seok Lee) **[정회원]**

- 1995년 2월 : 서울시립대학교 대학원 전자공학과 (공학석사)
- 1998년 2월 : 서울시립대학교 대학원 전자공학과 (공학박사)
- 1998년 3월 ~ 현재 : 청운대학교 인천캠퍼스 전자공학과 교수

〈관심분야〉 디지털신호처리, 뉴로모픽 시스템, 기계학습, 계산신경망