한국정보전자통신기술학회논문지(jkiiect)'23-2, Vol.16 No.1
논문 23-16-01-1 http://dx.doi.org/10.17661/jkiiect.2023.16.1.1

# 개체군 희소성 인덱스에 의한 컨벌루션 신경망 모델의 적대적 예제에 대한 뉴런 활동에 관한 연구

이영석*

# Study on Neuron Activities for Adversarial Examples in Convolutional Neural Network Model by Population Sparseness Index

Youngseok Lee*

**요 약** 시각 피질로부터 영감을 심층 신경망의 일종인 컨벌루션 신경망은 영상 관련 분야에서 이미 인간의 시각처리 능력을 넘어서 다양한 분야에 응용되고 있지만 적대적 공격의 출현으로 모델의 성능이 저하되는 심각한 위험에 노출되어 있다. 또한 적대적 공격에 대응하기 위한 방어 기술은 해당 공격에 효과를 보이지만 다른 종류의 공격에는 취약하다. 적대적 공격에 대응하기 위해서는 적대적 공격이 컨벌루션 신경망 내부에서 어떤 과정을 통하여 성능이 저하되는 지에 대한 분석이 필요하다. 본 연구에서는 신경생리학 분야에서 뉴런의 활동을 측정하기 위한 척도인 개체군 희소성 인덱스를 이용하여 AlexNet과 VGG11 모델의 적대적 공격에 대한 분석을 수행하였다. 수행된 연구를 통하여 적대적 예제에 대한 개체군 희소성 인덱스가 AlexNet에서는 전 연결 층에서 개체군 희소성이 증가하는 현상을 발견할 수 있었으며 이와 같은 동작은 일반적인 신경망의 동작에 반하는 결과로서 적대적 예제가 신경망의 동작에 영향을 미치고 있다는 강력한 증거이며 또한 동일한 실험을 실시한 VGG11에서는 전체 레이어에서 개체군 희소성 인덱스가 전반적으로 감소하여 개체 인식의 성능이 감소되는 활동을 관찰 할 수 있었다. 이와 같은 결과는 신경생리학적 관점에서 뉴런의 활동을 관찰하는 방식을 인공지능 분야에서도 활용하고 분석할 수 있는 방법을 제시하였다.

**Abstract** Convolutional neural networks have already been applied to various fields beyond human visual processing capabilities in the image processing area. However, they are exposed to a severe risk of deteriorating model performance due to the appearance of adversarial attacks. In addition, defense technology to respond to adversarial attacks is effective against the attack but is vulnerable to other types of attacks. Therefore, to respond to an adversarial attack, it is necessary to analyze how the performance of the adversarial attack deteriorates through the process inside the convolutional neural network. In this study, the adversarial attack of the Alexnet and VGG11 models was analyzed using the population sparseness index, a measure of neuronal activity in neurophysiology. Through the research, it was observed in each layer that the population sparsity index for adversarial examples showed differences from that of benign examples.

**Key Words** : Adversarial attack, CNN model, Neural activities, PSI, Visual cortex

## 1. 서론

시각 피질로부터 영감을 받은 심층 신경망의 일종인 컨벌루션 신경망(CNN: convolutional neural network)은 영상 분야에서 인간의 시각 처리 능력을 뛰어넘는 성과를 보이며 발전하고 있다[1]. 그러나 CNN은 인공지능의 보안과 관련된 적대적 공격(adversarial attack)으로 인하여 모델이 오 동작하는 문제점을 안고 있다. 이와 같은 문제를 해결하기 위해서는 CNN이 적대적 공격에 왜 취약한지에 대한 분석이 이루어져야 하지만 심층 신경망 내부는 비선형적인 연산 특성을 본질적으로 내포하고 있기 때문에 신

This paper is supported by Chungwoon University Research Fund in 2022.
*Corresponding Author : Dept. of Electronic Engineering, Chungwoon University (yslee@chungwoon.ac.kr)
Received January 20, 2023 Revised February 02, 2023 Accepted February 07, 2023

경망 내부에서 어떤 인과 관계에 의해 오동작하는 지에대한 적절한 설명이 없다. 이에 대한 해결 방법으로는 신경생리학 분야에서 연구하는 것과 같이 신경망의 각 레이어를 구성하고 있는 뉴런들의 활동을 분석하는 것이 필요하다.

본 연구에서는 신경생리학 분야에서 뉴런들의 활성화를 측정할 수 있는 척도인 PSI(Population Sparseness Index)[2]를 이용하여 정상적으로 동작하는 경우와 적대적 공격으로 인해 신경망 모델이 오동작하는 경우에 대하여 분석하였다. 또한 각 레이어에서 활성화되는 뉴런들의 활동 경향에 대한 분석을 수행하였다.

## 2. 관련 연구들

### 2.1 희소 코딩을 이용한 뉴런활동 분석

시각 피질을 구성하고 있는 뉴런들이 외부의 시각적 자극에 어떻게 반응하는지를 설명하기 위한 코딩이론은 하여 크게 세 가지로 나누어 설명된다 [3]. 첫 번째 이론은 외부의 자극에 대하여 뇌를 구성하는 모든 뉴런들이 반응한다는 것이고 [4], 두 번째 이론은 특정한 외부의 자극에 대하여 하나의 뉴런만 반응한다는 것이다[5]. 첫 번째 이론은 정보를 코딩하는데 필요한 만큼의 정보량을 제공할 수 있고 쉽게 일반화할 수 있으며 높은 강건성을 장점으로 하고 있고 두 번째 이론은 정보의 압축이나 정보를  인식하는데 필요한 에너지 손실을 최소화할 수 있다는 장점이 있다[5]. 최근에는 위의 두 이론의 장점만을 취하여 서로 다른 시각적 자극에 대하여 시각 피질을 담당하는 전체 뉴런들의 일부가 관여한다는 희소 코딩(sparse coding) 이론이 받아들여지고 있다[6]. 희소 코딩은 뇌의 신경망의 관점에서 PSI 라는 척도를 이용하여 측정되고 PSI의 값이 높으면 높을수록 해당되는 대상을 인식하는데 참여하는 뉴런의 수가 적다는 것을 의미한다[5]. 따라서 뇌를 자극하는 대상에 대하여 PSI를 측정하는 것은 해당 대상에 대한 뉴런의 활동을 간접적으로 관찰할 수 있는 측정 방법이다[8]. 최근 들어 뇌의 뉴런의 활동을 측정하기 위한 PSI를 인공 신경망에 응용하려는 시도가 있었다. [9]의 연구에서는 영상을 인식하는 CNN의 각 레이어에서 특정 객체의 인식에 참여하는 뉴런의 수를 계산하고 PSI를 추출함으로서 컨벌루션 신경망의 영상 분류 성능과 SPI 사이의 관계를 규명하였다. 또한 [7]과 [8]의 연구에서는 이와 같은 연구들은 해당되는 영상을 분류하기 위한 CNN 모델에 적용되어 CNN 모델을 구성하고 있는 각 레이어에서 약 10%-15%의 뉴런들만으로도 최고 성능의 약 90%에 해당하는 성능을 나타낸다고 보고하였다. 또한, 각 레이어에서 약 50%의 뉴런들을 제거한다고 해도 90%의 성능을 나타낸다고 주장하였다.

### 2.2 적대적 예제들 (adversarial examples)

CNN 모델은 사전 정의된 데이터셋을 심층 신경망(DNN: deep neural network)을 이용하여 학습하고, 학습된 모델에 학습에 사용되지 않은 데이터가 입력될 때에 학습한 모델의 파라미터들을 이용하여 학습된 정답에 가까운 출력을 생성하는 기계학습의 일종이다. 객체 인식 분야에서는 정상적인 동작의 경우 인간의 객체에 대한 인식률을 뛰어 넘는 성능을 발휘 할 수 있는 다양한 모델들이 연구되었다.

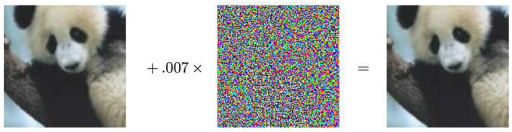
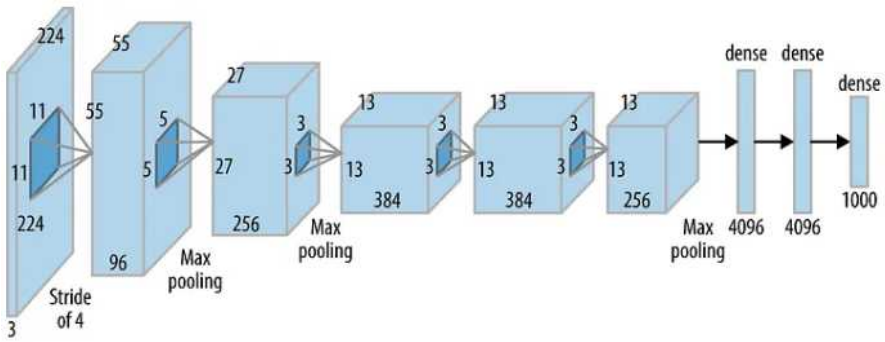
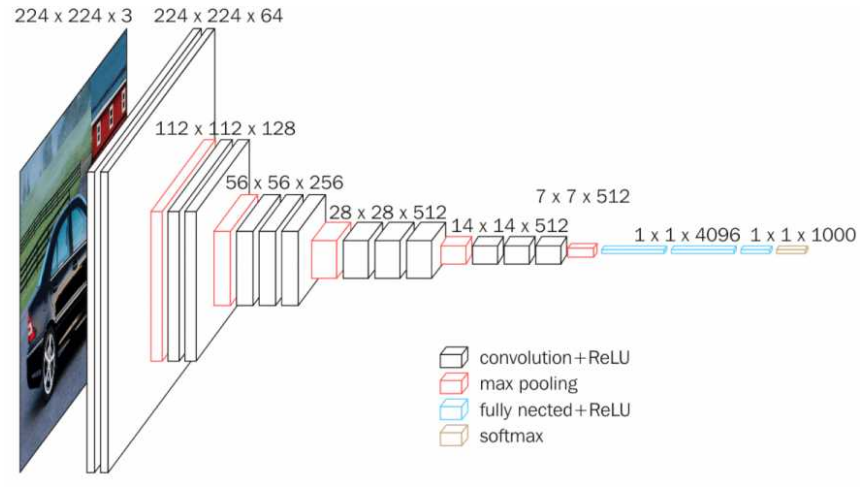



그림 1. 적대적 예제의 일례
Fig. 1. Example of adversarial example

그러나 육안으로 구분할 수 없는 작은 크기의 잡음을 추가하는 적대적 공격(adversarial attack)에 의해 신경망이 객체를 오 인식하는 적대적 공격 기술이 연구되면서 심층 신경망의 취약점이 나타났다. Szegedy[10] 등에 의하여 최초로 제안된 이 보안 공격 기술들은 CNN을 기반으로 하는 심층 신경망을 공격하여 정상적이 동작을 방해한다. 그림 1은 적대적 예제에 대한 일례로서 판다로 인식된 모델에 적절한 잡음을 추가하여 오 분류를 유도하는 공격의 일종인 FGSM(Fast Gradient Sign Method)를 보여 주고 있다. 적대적 예제의 대응하는 적대적 방어(adversarial defense) 기술들도 다양하게 연구되고 있고 어느 정




(a)




(b)

그림 2. 실험에 사용된 CNN 모델들 (a) AlexNet (b) VGG11.

Fig. 2. CNN models for experiment (a) AlexNet (b) VGG11.

도의 성과를 내고 있지만 해당 공격에는 강건한 성능을 발휘하지만 다른 방식으로 생성된 공격에는 취약한 특성을 나타낸다[11].

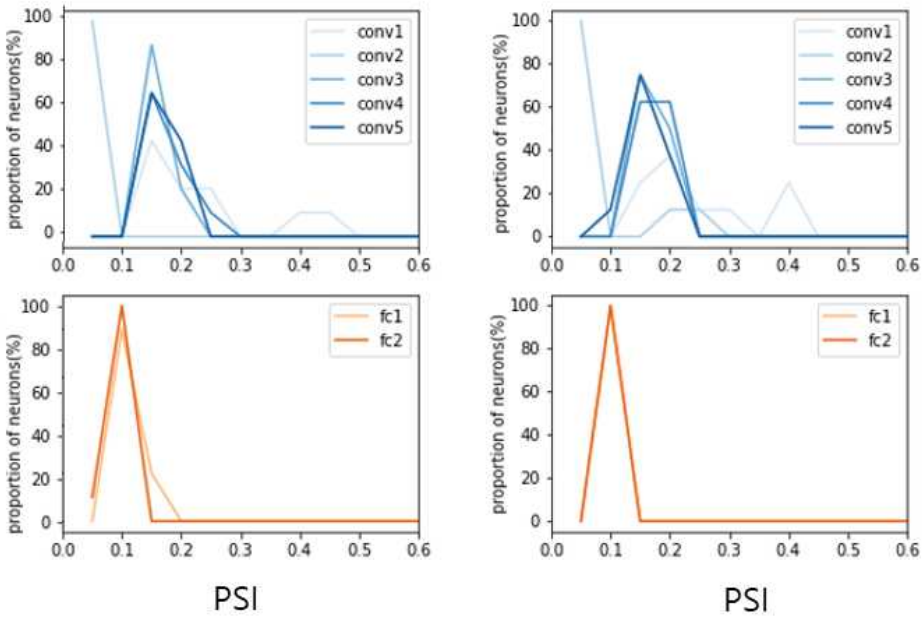
그러나 심층 신경망이 정상적으로 동작하는 상황에서도 심층 신경망 모델 내부에서 이루어지는 비선형 연산 등의 영향으로 어떠한 이유와 적절한 경로를 통하여 결과가 얻어지는 지에대한 뚜렷한 이론이 존재하지 않는다. 따라서 이와 적대적 공격을 분석하고 설명하기 위해서는 인간의 대뇌 피질 신경망의 관점에서 분석하고 설명할 수 있는 방법을 이용하여 인공 신경망을 분석하여 생물학적 신경망의 관점에서 인공 신경망을 해석하는 새로운 분석방법이 요구된다.

## 3. 실험 및 결과 고찰

### 3.1 CNN 모델과 영상 데이터셋

CNN 모델의 각 레이어의 영향을 미치는 뉴런들을 분석하기 위하여 그림 2에 나타낸 Alexnet[12]와 VGG11[13] 모델을 사용하였다. 두 모델들은 모두 파이토치로 구성된 사전 훈련된 모델을 사용하여 실험을 수행하였다. Alexnet은 5개의 컨벌루션 레이어와 3개의 전 결합(FC: fully connection) 레이어로 이루어져 있다. Alexnet의 컨벌루션 레이어 및 전 결합 레이어의 출력은 최종적으로 ReLU(Rectifying Linear Unit) 함수에 의하여 비선형적 특성을 지니고 있

VGG11은 Alexnet과 유사한 구조를 갖지만




(a) (b)

그림 3. AlexNet에서의 PSI와 범주별 인식 성능 그래프. (a) 온건한 예제 (b) 적대적 예제

Fig. 3. Graph of PSI vs. categories in Alexnet. (a) benign example (b) adversarial example

AlexNet이 $11\times11$의 수용장(receptive field)을 갖는데 비하여 더 작은 $3\times3$의 수용장을 갖고 있으며 컨벌루션 레이어의 수가 8개로 AlexNet 보다 많다. 실험을 위하여 사용된 영상 데이터셋은 5,000개의 학습 데이터와 10,000개의 검증 데이터로 이루어진 MNIST 데이터셋[13]을 이용하였다. 본 연구에서는 실험을 위하여 사전 훈련된 모델들을 사용하였기 때문에 MNIST 데이터셋을 검증 데이터를 사용하였다.

### 3.2 적대적 예제의 생성

적대적 예제를 생성하기 위한 적대적 공격 기술로서 식 (1)과 같은 FGSM[15]를 이용하였다.

$$x_{adv} = x + \delta,\ \| \delta \|_{\infty} \leq \epsilon \tag{1}$$

위 식에서 $x_{adv}$는 적대적 예제를 나타내고 $x$는 본 영상(original image), $\delta$는 적대적 공격을 위하여 생성된 잡음을 의미한다. 또한 $\epsilon$은 시각적 자극에 반응하지 않도록 눈에 보이지 않게 적대적 예제를 성하기 위한 잡음 $\delta$의 크기로서 경험적인 방법에 의하여 만들 수 있으며 본 연구에서는 0.03을 경험적으로 구하여 적용하였다.

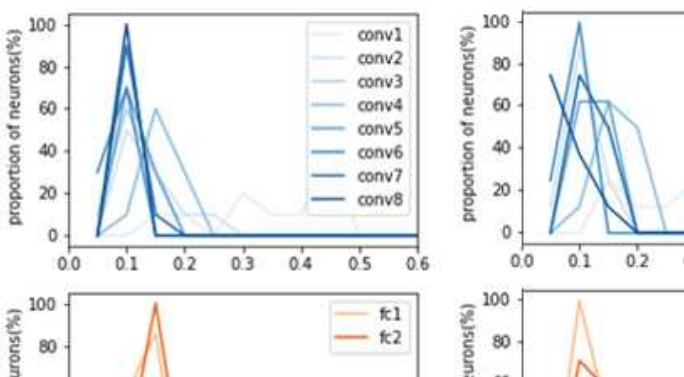
$$\delta = \epsilon sign(\nabla_x L(\theta, x, y)) \tag{2}$$

적절한 잡음을 찾기 위한 $\delta$는 식(2)와같이 적절한 비용함수 $L$이 주어지는 경우 CNN 모델의 파라미터 $\theta$를 이용하여 정답 레이블 $y$의 반대 방향으로 기울기(gradient)를 갱신하여 생성할 수 있다.

### 3.3 Population sparseness index(PSI)

두 모델의 각 레이어에서 뉴런의 활동을 추출하기 위한 PSI는 식 (3)과 같이 정의된다.

$$PSI = \frac{1-a}{1-\frac{1}{N_u}} \tag{3}$$



위 식에서 $N_u$는 각 레이어에 있는 뉴런들의 수이고, $a$는 각 레이어에 있는 뉴런들의 수와 객체 인식을 위하여 해당 범주에 대응하여 활성화되는 뉴런들의 수에 대한 비율로서 식 (4)와 같이 정의된다[16].

$$a = \frac{(\sum r_u / N_u)^2}{\sum (r_u)^2 / N_u} \tag{4}$$

위 식에서 $r_u$는 해당 레이어에서 활성화되는 뉴런의 수를 나타낸다.

### 3.4 실험 결과

Alexnet에서 추출된 PSI를 이용한 실험 결과는 그림 3과 같이 카테고리 별 성능과 PSI의 그래프로 나타내었다. 그림 3(a)는 MNIST 데이터셋에서 제공하는 검증 데이터셋의 결과이고 그림 3(b)는 FGSM 공격을 받은 적대적 예제에 대한 결과이다. 5개의 컨벌루션 레이어에서의 PSI 분석에서 1번 레이어의 PSI 값이 적대적 예제의 경우 더 큰 값을 나타내고 PSI의 분포가 피크 값을 갖는 것을 관찰할 수 있다.

또한 4번 컨벌루션 레이어의 PSI의 값이 온건한 예제와 적대적 예제에서 0.65 근처에서 유사한 값을 갖지만 인식률에서는 온건한 예제는 60.2%, 적대적 예제에서는 63.4%의 성능을 나타내는 것을 확인 할 수 있다.

표 1. AlexNet의 각 레이어별 PSI 피크 값 분포
Table 1. Distribution of PSI-peak values in each layer of AlexNet

| Distribution of peak values in layers of VHH11 | | |
|---|---|---|
| Layer | Benign dataset | Attacked dataset |
| conv1 | 39.4 | 68.2 |
| conv2 | 93.4 | 82.1 |
| conv3 | 84.2 | 59.2 |
| conv4 | 60.2 | 63.4 |
| conv5 | 60.3 | 63.4 |
| fc1 | 79.6 | 99.8 |
| fc2 | 94.2 | 99.4 |

표 1은 각 레이어에서 피크 값을 갖는 PSI값들을 표시한 것으로 이와 같은 결과는 적대적 예제의 경우 공격으로 인하여 올바른 레이블을 찾지 못하여 레이어에서 적절한 결과를 내지 못하는 것으로 추론할 수 있다. 이와 같은 결과는 전형적인 CNN 신경망에서 전 연결의 뉴런들이 각 레이블에 해당하는 뉴런들에 의하여 희소성이 증가하는 경향에 반하는 결과로 해석할 수 있으며 비록 전 연결 층의 희소성이 감소한 결과는 적절한 추론 과정을 통하여 올바른 레이블링을 하지 못하는 비정상적 동작으로 추정할 수 있다.

또한 전 연결 층에서 PVGG11 모델에 대하여 동일한 실험을 한 결과는 그림 4에 나타내었다. VGG11 모델의 실험 결과는 Alexnet의 실험 결과를 더 명확히 하여 적대적 예제에 의한 오동작과 PSI 사이의 관계를 더욱 잘 나타낸다. 총 8개의 컨벌루션 레이어 가운데 최종단의 8번 레이어는 PSI의 값이 거의 0.05 근방에서 형성되는 것을 관찰 할 수 있으며 이와 같은 결과는 8번 컨벌루션 레이어에서 거의 모든 뉴런들이 모델의 성능에 영향을 미치는 것으로 분석할 수 있다. 표 2는 VGG11에서 각 레이이별 PSI의 피크 값을 나타낸 것으로 PSI가 레이어가 진해됨에 따라 높은 우 편향 경향을 관찰할 수 있는데 비하여 공격을 받은 데이터셋의 결과는 느리게 우 편향되거나 변화하지 않는 경향이 있음을 관찰 할 수 있다.

표 2. VGG11의 각 레이어별 PSI 피크 값 분포
Table 2. Distribution of PSI-peak values in each layer of VGG11

| Distribution of peak values in layers of VGG11 | | |
|---|---|---|
| Layer | Benign dataset | Attacked dataset |
| conv1 | 18.6 | 28.8 |
| conv2 | 11.2 | 78.3 |
| conv3 | 47.3 | 36.3 |
| conv4 | 58.2 | 36.6 |
| conv5 | 77.6 | 36.5 |
| conv6 | 77.5 | 80.7 |
| conv7 | 98.3 | 60.5 |
| conv8 | 99.8 | 61.2 |
| fc1 | 60.1 | 70.2 |
| fc2 | 78.4 | 51.4 |

또한 8번 컨벌루션 레이어에서 온전한 샘플은 98%, 2번 전 연결 층에서 약 78%의 성능을 나타는 반면에 적대적 예제의 경우 는 약 60%의 성능을 나타내었다. 전 연결 층의 경우 2개의 연결 층에서 모두 차이를 나타낸 것을 관찰할 수 있고 특히 적대적 예제의 경우 약 50%의 낮은 성능을 보이는 것을 확인할 수 있다. 또한 PSI 비교에서 온건한 샘플의 경우 0.15를 나타낸데 비하여 적대적 예제의 경우 0.1에서 피크 값을 형성한 것을 확인할 수 있다. 이와 같은 결과는 적대적 예제가 CNN 모델에 적용되는 경우 낮은 성능과 함께 낮은 PSI 값을 나타낸다는 것을 의미한다.

## 4. 결론

본 연구에서는 신경생리학 분야에서 뉴런의 활동을 관찰하기 위하여 사용한 PSI 분석을 통하여 대표적인 CNN 모델인 Alexnet과 VGG11에 대한 각 레이어별 성능과 뉴런의 활동에 대하여 관찰하는 실험 및 분석을 수행하였다.

두 모델에 대한 관찰로부터 정상적인 MNIST 데이터셋에 대한 결과와 적대적 예제에 대한 결과가 PSI의 관점에서 서로 다른 결과를 나타낼 뿐만 아니라 이 결과의 차이가 성능과 연결될 수 있는 가능성을 보여준다. 이와 같은 결과는 심층신경망에서 적대적 예제를 뇌를 기반으로 하는 연구들과 관련하여 분석할 수 있는 가능성을 제시하였다.

## REFERENCES

[1] Yuan, Xiaoyong, "Adversarial examples: Attacks and defenses for deep learning", IEEE transactions on neural networks and learning systems, Vol. 39, No. 9, pp. 2805-2824, 2019.

[2] Quian Quiroga, Rodrigo, and Gabriel Kreiman, "Measuring sparseness in the brain," comment on Bowers, Vol. 291, 2009.

[3] Szegedy Christian, "Going deeper with convolutions", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.

[4] Agrawal Pulkit, Girishick Ross, "Analyzing the performance of multilayer neural networks for object recognition", European conference on computer vision, pp. 329-344, 2014.

[5] Li I, Yixuan, "Convergent learning: Do different neural networks learn the same representations?", arXiv preprint arXiv:1511.07543, 2015.

[6] Wang, Jianyu, "Unsupervised learning of object semantic parts from internal states of CNNs by population encoding", arXiv-1511.arXiv pe-prints: arXiv-1511. 2015.

[7] Morcos, Ari, "On the importance of single directions for generalization", arXiv preprint arXiv:1803.06959, 2018.

[8] Casper, Stephen, "Removable and/or repeated units emerge in overparametrized deep neural networks", ArXiv abs/1912.04783, 2019.

[9] Tavanaei, Amirhossein, Anthony S. Maida, "Bio-inspired spiking convolutional neural network using layer-wise sparse coding and STDP learning", arXiv preprint arXiv:1611.03000, 2016.

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

[11] Qiu, Shilin, "Review of artificial intelligence adversarial attack and defense technologies", Applied Sciences, Vol. 9, No. 5, 2019.

[12] Alom, Zahangir, "The history began from alexnet: A comprehensive survey on deep learning approaches", A comprehensive survey on deep learning approaches arXiv preprint arXiv: 1803.01164, 2018.

[13] Gan, Yanfen, Jixiang Yang, and Wenda Lai, "Video object forgery detection algorithm based on VGG-11 convolutional neural network", 2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS). IEEE, 2019.

[14] https://sdc-james.gitbook.io/onebook/4.-an

d/5.1./5.1.3.mnist-dataset.
[15] Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572, 2014.
[16] Vinje William, and Jack Gallant. "Sparse coding and decorrelation in primary visual cortex during natural vision", Science 287.5456, pp.1273-1276 2000.

## 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, 임베디드시스템, 기계학습, 계산신경망