https://doi.org/10.22974/jkda.2022.60.5.005

투고일 : 2022. 4. 8 심사일 : 2022. 4. 18 게재확정일 : 2022. 4. 30

인공지능 딥러닝의 역사와 현황, 그리고 미래 방향

이원진

서울대학교 치의학대학원 영상치의학교실

ORCID ID

Won-Jin Yi, MS, PhD. Professor, https://orcid.org/0000-0002-5977-6634

ABSTRACT

History, Current Status and Future Directions of Deep Learning

Won-Jin Yi

Department of Oral and Maxillofacial Radiology, School of Dentistry, Seoul National University

Deep learning is a subset of machine learning, and machine learning is also a subset of artificial intelligence (AI). The biggest difference between machine learning and deep learning is that in the learning of artificial intelligence models, machine learning basically requires a human feature extraction process before learning, but deep learning does not require this process and the original data is directly used as input. The development of deep learning coincides with the development of artificial neural networks (ANNs), and many people have contributed to the development of artificial neural networks for decades. The following five models are the representative architectures most widely used in deep learning. That is, Deep Feedforward Neural Network (D-FFNN), Convolutional Neural Network (CNN), Deep Belief Network (DBN), Autoencoders (AE), and Long Short-Term Memory (LSTM) Network. A convolutional neural network (CNN) is a feedforward NN composed of a convolutional layer, a ReLU activation function, and a pooling layer. CNNs provide properties of weight sharing and local connectivity to process high-dimensional data. In dental and medical fields, an AI model that can be interpretable or explainable (XAI) is needed to increase patient persuasiveness. In the future, explainable AI (XAI) will become an indispensable and practical component in order to obtain an improved, transparent, secure, fair and unbiased AI learning model.

Key words : Artificial Intelligence (AI), Deep learning, Artificial Neural Networks (ANN), Convolutional Neural Network (CNN), Explainable AI (XAI)

Corresponding Author
Won-Jin Yi, MS, PhD. Professor
Department of Oral and Maxillofacial Radiology, School of Dentistry, Seoul National University, 101, Daehak-ro, Jongno-gu, Seoul, 03080, Korea
Tel : 82-2-2072-3049 / Fax : 82-2-741-0401 / E-mail : wjyi@snu.ac.kr

1. 딥러닝(deep learning)의 역사

최근 딥러닝(deep learning)이 영상분석 및 음성인식 분야에서 획기적인 성능을 보여주고 있다. 딥러닝은 이러한 복잡한 예측 모델을 학습하는 데 사용할 수 있으며, 단일 방법이 아니라 학습 알고리즘 군을 말한다. 딥러닝(deep learning)은 머신러닝(machine learning)의 한 부분집합이며, 또한 머신러닝은 인공지능(Artificial Intelligence, AI)의 한 부분집합이다(그림 1). 머신러닝은 컴퓨터를 이용하여 입력 데이터에 대해 다른 공간에서 더 나은 표현(Representation, Mapping)을 찾는 자동화된 과정이다. 기존의 머신러닝과 딥러닝의 가장 큰 차이는 인공지능 모델의 학습에 있어서, 머신러닝은 기본적으로 모델의 학습 전에 인간에 의한 특징추출(feature extraction) 과정이 필요하나, 딥러닝은 이러한 과정이 필요하지 않고 원래 데이터가 바로 모델의 입력으로 사용된다(그림2). 또 하나의 큰 차이점은 머신러닝은 학습 데이터의 수가

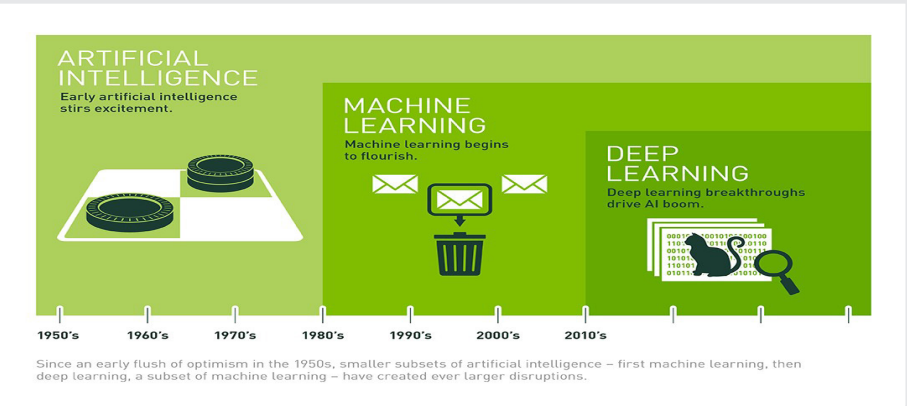

그림 1. 인공지능, 머신러닝 및 딥러닝의 관계

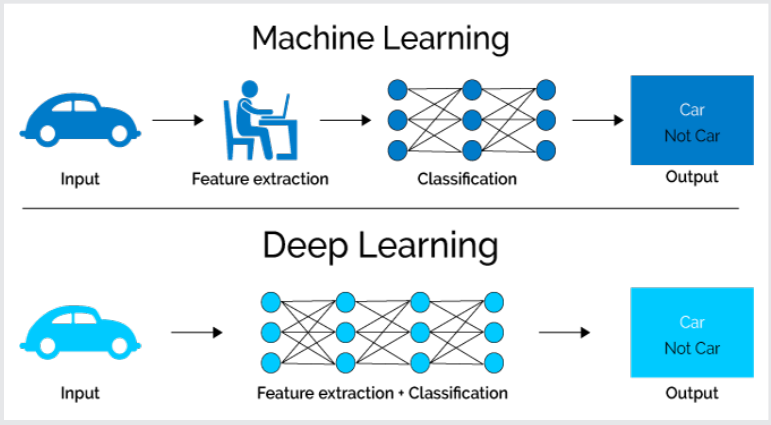

그림 2. 머신러닝과 딥러닝의 차이점

계속 증가하여도 일정 단계에 도달하면 모델의 성능이 포화단계에 도달하여 더 이상 향상되지 않지만, 딥러닝은 데이터 수가 증가하면 할수록 성능이 계속 향상되는 특징을 가진다(그림 3). 그리고 인간이 반복적인 경험과 인식을 통해 학습하는 방식과 매우 유사한 방식으로, 딥러닝에서는 인공지능 모델을 학습시킬 수 있다. 즉 어린이에게 많은 사진과 실제 애완동물을 보여 주며 개와 고양이를 인식하는 방법을 가르치는 것처럼, 딥러닝 모델에서도 많은 데이터를 제공하여 영상과 패턴을 인식하는 방법을 학습시킬 수 있다.

딥러닝의 발전은 인공신경망(Artificial Neural Network, ANN)의 발전과 그 궤를 같이하여, 많은 사람들이 수십 년 동안 인공신경망의 발전에 기여했다. 딥러닝 발전의 역사를 간략히 정리하면 다음과 같다. 1943년, McCulloch와 Pitts가 처음으로 뉴런의 수학적 모델을 만들었으며, 이 모델은 실제 생물학적 뉴런의 생물학적 메커니즘을 모방하지 않고, 뉴런의 기능에 대한 추상적인 공식을 제공하는 것을 목표로 했고 이 모델은 학습을 고려하지 않았다(그림 4)[1]. McCulloch-Pitts 뉴런은 입력을 받고 가중치 합을 취하여 결과가 임계값 미만이면 '0'을 출력하고 그렇지 않으면 '1'을 출력한다. 1949년, Hebb에 의해서 생물학에서 유추된 신경망 학습 방법이 처음으로 소개됐다[2]. Hebbian 학습 방법은 비지도 신경망 학습(unsupervised learning) 방법의 한 형태이며, 뉴런의 가소성에 근거하여 뉴런을 계속적으로 자주 사용함에 따라서 그 뉴런 간의 경로가 더 강화된다. 1957년, Rosenblatt에 의해 최초로 학습할 수 있는 신경망인 Mark I Perceptron이 제안됐다[3]. Mark I Perceptron은 선형 이진 분류기 역할을 하는 단일 층(layer) 신경망이다. Perceptron의 뉴런은 바이어스(bias)인 시냅스 가중치(b)를 추가적으로 입력 받고, Heaviside 함수를 활성화 함수(activation function)로 사용했다(그림 5). Mark I Perceptron의 장점은 원하는 출력과 실제 출

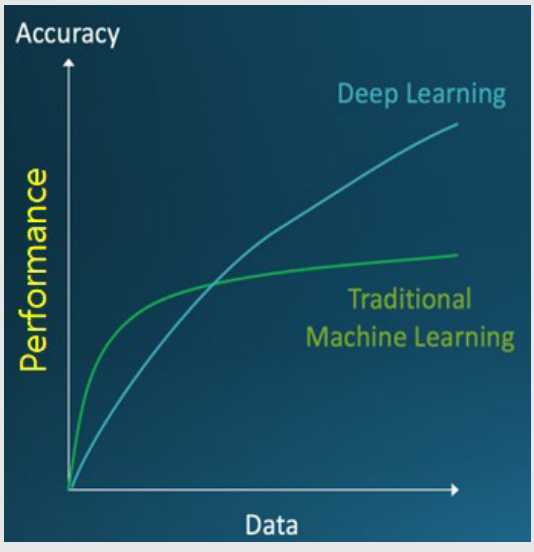

그림 3. 학습데이터 증가에 따른 머신러닝과 딥러닝의 성능 향상 비교

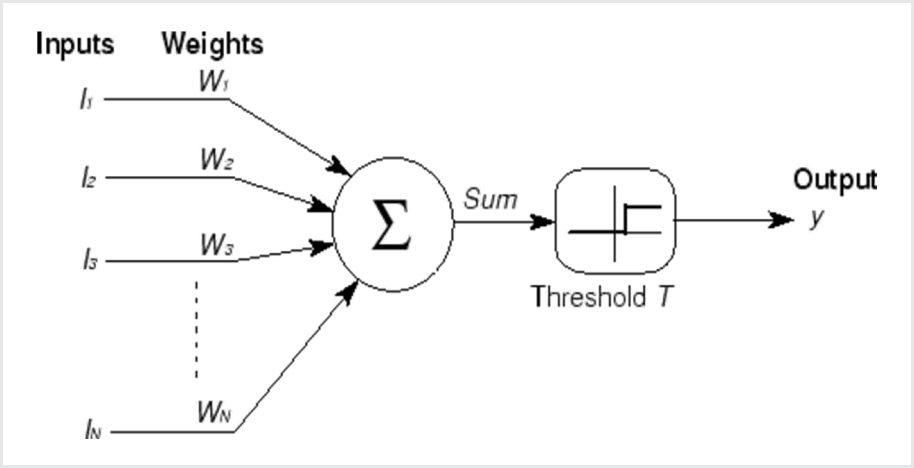

그림 4. McCulloch와 Pitts에 의해 제안된 최초의 뉴런의 컴퓨터 모델

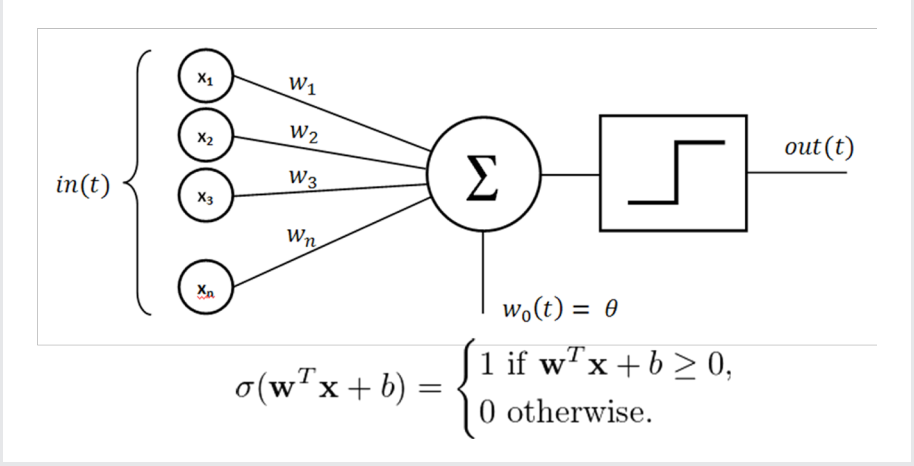

그림 5. Frank Rosenblatt에 의해 제안된 Mark I Perceptron

력의 차이를 최소화하면서 연속적으로 전달된 입력을 통해 가중치를 학습할 수 있었다. 그러나 이는 선형으로 분리 가능한 클래스만 분리하는 방법만 학습할 수 있어서, 단순한 비선형 함수나 XOR(exclusive-OR)는 학습할 수 없는 한계를 보였다(그림 6).

1960년, Perceptron 학습을 위한 델타(Delta) 학습규칙이 Widrow와 Hoff에 의해 소개됐다[4]. Widrow & Hoff 학습규칙 또는 최소자승법(Least Mean Square)

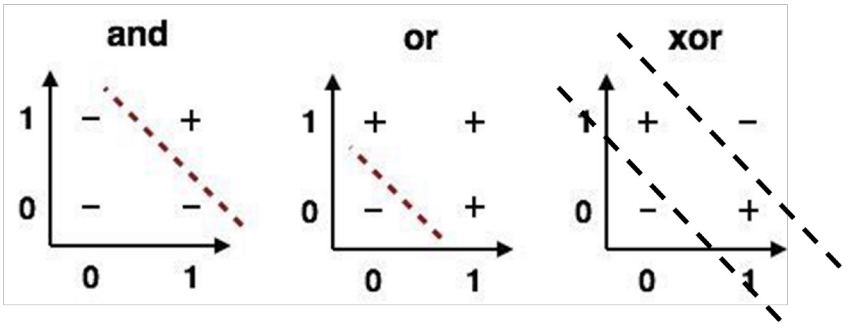

그림 6. Mark I Perceptron은 XOR(exclusive-OR)는 학습할 수 없음

학습규칙 이라고도 하는 델타 학습규칙은 뉴런의 가중치를 업데이트하기 위한 경사하강법(gradient descent)에 기반한 학습규칙이다. 1968년, Ivakhnenko는 신경망 훈련을 위한 GMDH(Group Method of Data Handling)라는 방법을 도입했다[5]. 이는 순방향 다층 퍼셉트론(Feedforward Multilayer Perceptron) 유형의 첫 번째 딥러닝 네트워크로 간주된다. 이를 바탕으로 1971년, Ivakhnenko는 8개의 층이 있는 심층 GMDH 네트워크를 사용했으며, 흥미롭게도 층의 수와 층당 단위는 처음부터 고정되지 않고 학습될 수 있었다[6]. 1969년, Minsky와 Papert의 중요한 논문이 발표되었는데, XOR 문제는 선형으로 분리할 수 없기 때문에 한 개의 층으로 이루어진 Perceptron으로 학습할 수 없다는 것을 보여준다[7]. 이것은 첫 번째 AI 겨울(the first AI winter)이라고 불리는 인공지능 암흑기를 촉발하게 된다.

1974년, 지도학습(supervised learning) 방식으로 가중치를 학습하기 위해 오류역전파(Error back-propagation) 알고리즘이 제안되고[8], 신경망에 적용됐다[9]. 1980년, Fukushima에 의해 Neocognitron이라는 시각적 패턴 인식을 위한 다층 신경망이 도입됐다(그림 7)[10]. 심층 GMDH 네트워크 다음으로 Neocognitron 인공신경망은 두 번째 심층 신경망으로 인정된다. 여기서, 처음으로 콘볼류션 신경망(Convolutional Neural Network, CNN)이 도입됐다. Neocognitron은 현재의 지도학습 기반 심층 순방향 신경망(Deep Feedforward NN, D-FFNN)의 아키텍처와 매우 유사하다[11]. 1982년, Hopfield는 현재 Hopfield Network이라고 불리는 내용 주소 기억장치(content address memory) 신경망을 도입했다[12]. Hopfield Network는 순환 신경망(recurrent neural network, RNN)의 하나이다. 1986년, Rumelhart 등의 논문에서 오류역전파 알고리즘이 다시 등장했다[13]. 그들은 오류역전파 학습 알고리즘이 내부 표현을 유용하게 생성할 수 있고, 따라서 일반적인 신경망 학습 작업에 사용할 수 있음을 실험적으로 보여주었다[13]. 다층 신경망 아키텍처(Multilayer perceptron)에 오류역전파 알고리즘을 적용한 것은 인공지능 및 인지과학에서 중요한 돌파구가 됐으며, 딥러닝 연구의 새로운 장을 열었다(그림 8). 1989년, 필기체 숫자를 학습하기 위해서, CNN이 오

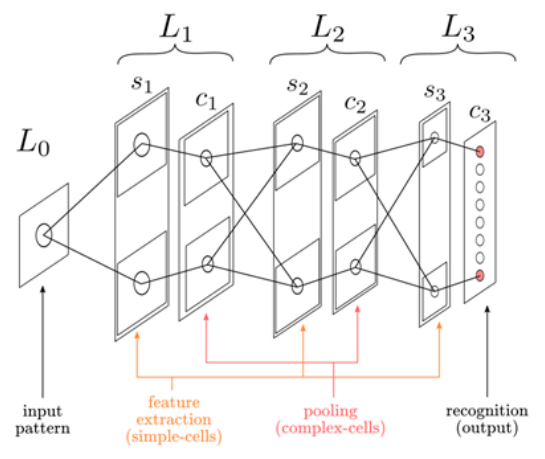

그림 7. Fukushima에 의해 개발된 Neocognitron 신경망 구조

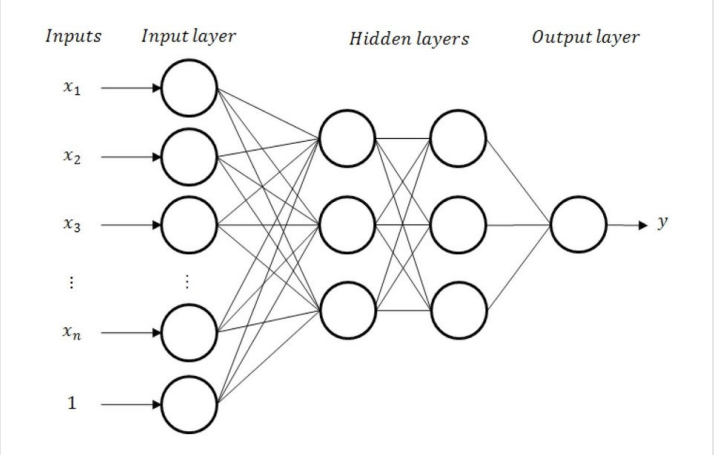

그림 8. 다층 신경망 아키텍처(Multilayer perceptron)

류역전파 알고리즘을 이용하여 학습됐다[14]. 이 CNN의 발전된 버전이 나중에 90년대 후반과 2000년대 초반에 미국에서 수표와 우편번호를 읽는 데 사용됐으며, 필기체 숫자 인식에서 99.05%의 정확도를 보였다(그림 9) [15, 16]. 1991년, Hochreiter는 오류역전파 알고리즘으로 훈련할 수 없는 문제와 관련된 모든 딥러닝 네트워크의 근본적인 문제를 연구했다[17]. 그의 연구에 따르면 오류역전파 알고리즘에서 역전되는 오류

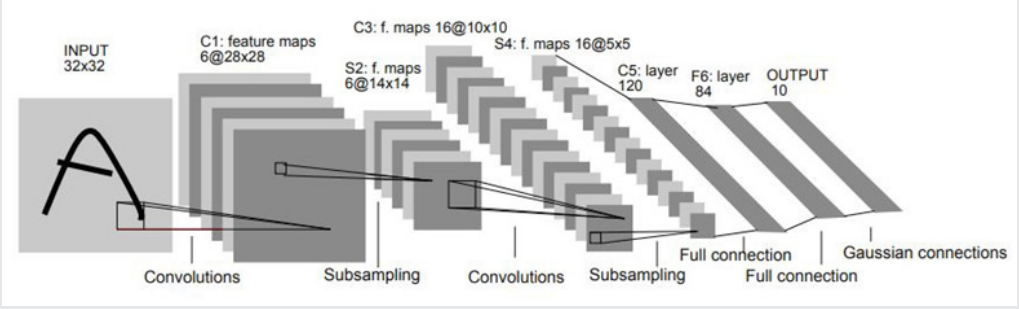

그림 9. 필기체 숫자 인식에 활용된 CNN 일종인 LeNet5 신경망 모델 구조

는 제한 없이 감소하거나 증가할 수 있고, 감소의 경우 네트워크 깊이에 비례한다. 이것은 심층 신경망 학습에서 기울기 소실(gradient vanishing)과 기울기 폭발(gradient exploding) 문제로 알려져 있다. 이로 인해서 심층 신경망의 경우 오류역전파 알고리즘으로 충분히 학습할 수 없음이 알려진다. 이것은 두 번째 AI 겨울(the second AI winter)이라고 불리는 인공지능 암흑기를 촉발한다.

1992년, 기울기 소실(gradient vanishing)과 기울기 폭발(gradient exploding) 문제에 대한 첫 번째 부분적인 해결책이 Schmidhuber에 의해 제안됐다[18]. 이 아이디어는 후속 지도학습을 가속화하기 위해 비지도 학습 방법으로 순환 신경망(RNN)을 사전 훈련시키는 것이다. 이때 사용된 RNN은 1,000개 이상의 층을 가졌다. 1997년, Hochreiter와 Schmidhuber에 의해 RNN 학습을 위한 최초의 지도학습 모델이 소개되었으며, 이를 LSTM(Long Short-Term Memory)이라고 한다[19]. LSTM은 신경망이 더 오랜 기간 동안 정보를 기억하게 하여 층 간의 기울기 소실 문제를 방지한다. 1998년, CNN의 학습을 향상시키기 위해 오류역전파 알고리즘과 확률적 경사하강법(Stochastic Gradient Descent algorithm) 알고리즘이 결합된다[16]. 그 결과 수표에서 필기체 숫자를 인식하기 위해 7층의 CNN LeNet5가 도입됐다(그림 9). 2006년, Deep Belief Networks라고 하는 신경망은 "Greedy Layer-Wise Training"이라는 전략을 사용하여 심층 신경망이 효율적으로 학습될 수 있음을 보여준다[20]. 이는 딥러닝이라는 용어의 사용을 대중화하고 제3의 신경망 물결이 시작되는 계기가 됐다. 2012년, Alex Krizhevsky는 GPU를 활용하고 LeNet5를 개선한 CNN인 AlexNet을 사용하여 ImageNet Large Scale Visual Recognition Challenge(ImageNet LSVRC)에서 우승했다(그림 10)[21, 22]. ILSVRC는 백만 개 이상의 영상 데이터를 1000개 이상의 클래스로 분류하는 대회이며, 이전까지 영상분류 오차가 20%이상이었으나 AlexNet이 15%의 오차를 달성했다. AlexNet의 성공으로 딥러닝에서 새로운 돌파구가 열리면서 CNN의 르네상스가 시작됐다. 그 후에, 2014년에 Goodfellow 등이 적대적 생성 신경망(Generative Adversarial Networks, GAN)을 발표했다(그림 11)[23]. GAN은 생성과 판별하는 두 개의 신경망이 게임과 같은 방식으로 서로 적대적으로 경쟁하면서, 입력과 다른 새로운 데이터를 생성할 수 있다. 이는 Yann LeCun에 의해 "지난 20년 동안 가장 멋진 기계학습 아이디어"라고 불렸다.

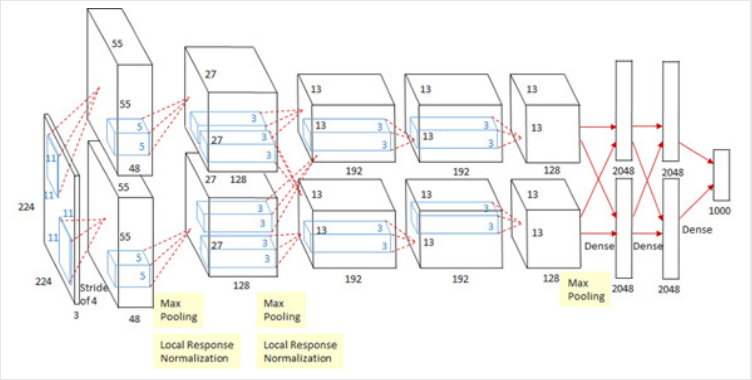

그림 10. CNN 일종인 AlexNet 신경망 모델 구조

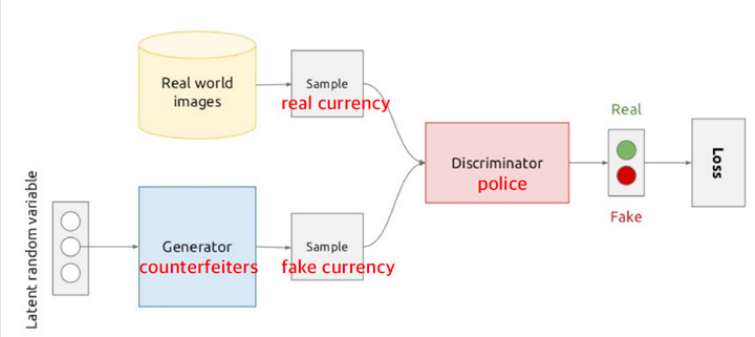

그림 11. 적대적 생성 신경망(Generative Adversarial Networks, GAN) 구조

그림 12. 컴퓨팅 분야에서 노벨상으로 불리는 튜링상(ACM A.M. Turing Award) 2019년 수상자

2019년, Yoshua Bengio, Geoffrey Hinton과 Yann LeCun은 심층 신경망을 컴퓨팅의 중요한 구성요소로 만든 개념과 엔지니어링의 혁신에 대한 공로를 인정받아 컴퓨팅 분야에서 노벨상으로 불리는 튜링상(ACM A.M. Turing Award)을 수상했다(그림 12).

2. 딥러닝(deep learning)의 현황

Gartner사에서 2019년부터 해마다 인공지능 기술에 대한 Hype Cycle을 발표하고 있으며, 2021년의 인공지능 기술의 전반적인 트렌드는 다음과 같다(그림13). 즉, 기업 운용에서 AI 솔류션의 효율적 활용, 효율적인 데이터/모델과 컴퓨팅 파워의 활용을 위한 혁신, 개인정보 보호 및 사회안전 도모를 위한 공정하고 투명한 AI 시스템, 그리고 적은 데이터와 넓은 데이터를 활용한 기술 개발이다. 현재 딥러닝에서 가장 많이 활용되고 있는 대표적인 아키텍처는 다음 5개의 모델이다. 즉 Deep Feedforward Neural Network(D-FFNN), Convolutional Neural Network(CNN), Deep Belief Network(DBN), Autoencoders(AE), 및 Long Short-Term Memory(LSTM) Network이다(그림 14). 모든 딥러닝 모델의 공통적인 특성은 표현학습(representation learning)을 수행한다는 것이다. 이것은 특징학습(feature learning)이라고도 하며, 입력 데이터에 대해서 새롭고 더 나은 표현을 자동적으로 학습하는 것을 의미한다. 여기서 딥러닝 모델은 한 단계 내에서 최종 표현을 학습하는 것이 아니라, 다수의 은닉층(hidden layer) 간의 다 단계 표현 변환을 통하여 여러 단계에 걸쳐서 특징을 자동적으로 학습한다[15]. 딥러닝 모델의 또 다른 일반적인 특성은 층 간의 변환이 비선형이라는 것이며, 이는 모델의 표현력을 증가시킨다[24]. 또한 개별 표현은 수동으로 설계되지 않고 학습 데이터를 통해 자동적으로 학습되며[15], 이는 딥러닝 모델을 여러 응용에서 매우 유연하게 만든다.

1) 심층 순방향 신경망 (Deep Feedforward Neural Network, D-FFNN)

가장 일반적인 딥러닝 구조는 심층 순방향 신경망(Deep Feedforward NN, D-FFNN)에 해당하는 완전 연결된 순방향 신경망이다. 하나의 은닉층(hid-

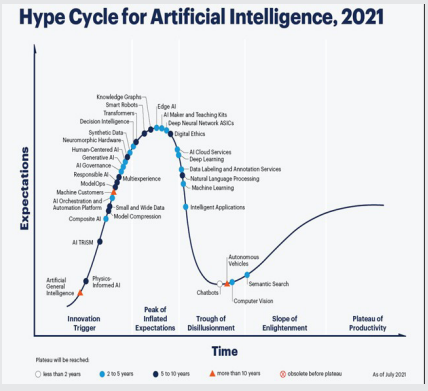

그림13. Gartner 사에서 발표한 인공지능 기술에 대한 Hype Cycle(2021)

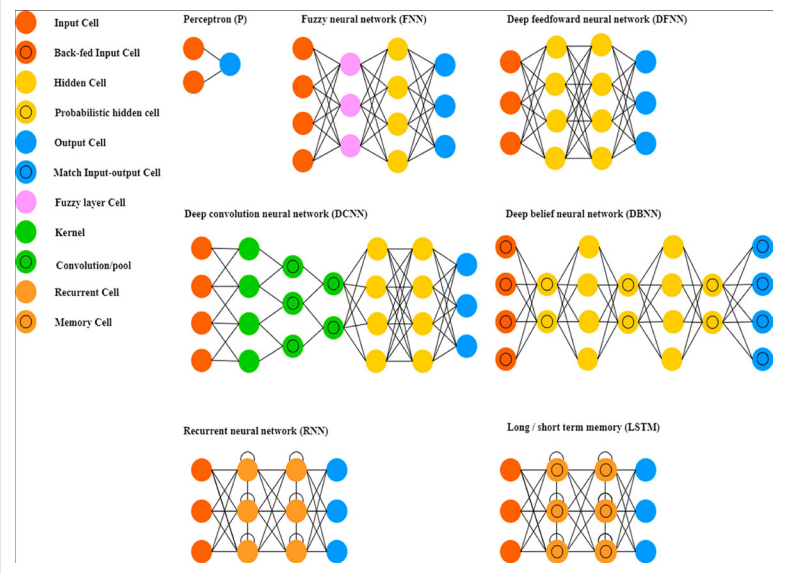

그림 14. 대표적인 인공지능 딥러닝 모델 구조

den layer)과 유한한 수의 뉴런이 있는 순방향 신경망은 모든 연속함수를 근사할 수 있음이 증명됐으며[25], 이것을 보편적 근사정리(universal approximation theorem)라고 한다. 이때, 하나 이상의 은닉층이 있는 FFNN을 사용하지 않았던 이유는, 보편적 근사정리가 그러한 네트워크를 학습하는 방법에 대한 정보를 제공하지 않기 때문에 매우 어려운 것으로 판명됐기 때문이다. 하나의 은닉층을 가지는 네트워크 학습의 어려움은 은닉층의 너비가 기하급수적으로 커질 수 있다는 것이다. 그러나, 많은 은닉층과 제한된 수의 뉴런을 가진 FFNN에서 학습 알고리즘이 발견되고, 보편적 근사정리는 입증될 수 있었다[26]. 따라서 심층 순방향 신경망(D-FFNN)이 보편적으로 사용될 수 있게 되었다. D-FFNN의 매개변수에 대한 실제 학습은 오류 역전파 알고리즘을 사용하여 수행될 수 있으며, 현재는 계산 효율성을 위해 확률적 경사 하강법(Stochastic Gradient Descent)이 사용된다[27]. 확률적 경사 하강법은 무작위로 선택된 훈련 배치 세트(batch)에 대한 경사를 계산하고, 이 배치 세트에 대한 매개변수를 순차적으로 업데이트한다. 결과적으로, 더 빠른 학습이 가능하지만, 정밀도가 감소한다는 단점이 있다. 그러나 많은 수의 표본(빅 데이터)이 있는 데이터 세트의 경우, 빠른 학습의 장점이 단점보다 더 많다. 흥미롭게도, Mayr의 연구에서 D-FFNN이 약물의 독성 예측에서 다른 방법을 능가하는 것으로 나타났다[28]. 예제로서, 약물 표적 예측에서 D-FFNN이 다른 방법에 비해 우수한 것으로 나타났다[29]. 이는 D-FFNN 아키텍처도 현대의 응용 프로그램에서 성공적으로 사용될 수 있음을 보여줬다.

2) 컨볼루션 신경망 (Convolutional Neural Network, CNN)

콘볼류션 신경망(Convolutional Neural Network,

CNN)은 컨볼루션층(convolutional layer), ReLU 활성화 함수(activation function) 및 풀링층(pooling layer)으로 구성되는 순방향 신경망(Feedforward NN)이다. 심층 CNN(deep CNN)은 일반적으로 컨볼루션, 풀링 및 완전 연결(fully connected layer) 층을 포함한 다수의 순방향 신경망 층으로 구성된다(그림 15). 일반적으로 기존 ANN에서 한 층의 각 뉴런은 다음 층의 모든 뉴런에 연결되고, 각 연결은 네트워크의 학습을 통해 결정된다. 이로 인해 매우 많은 수의 매개변수(weights)가 발생한다. 반면에, CNN은 컨볼루션(convolutional) 연산을 통해서 모두 연결된 층을 사용하는 대신, 뉴런 간의 로컬 연결(local connectivity)을 사용한다. 즉, 컨볼루션층(convolutional layer)에서 뉴런은 다음 층의 근처 뉴런에만 연결된다. 이렇게 되면 네트워크의 총 매개변수 수를 크게 줄일 수 있다. 또한, 여기서, 국부적 수용영역(local receptive fields)과 뉴런 사이의 모든 연결은 동일한 가중치 셋(set of weights)을 사용하며, 이 가중치 셋을 커널(kernel)이라고 한다. 이 커널은 로컬 수용영역에 연결된 다른 모든 뉴런에 동일하게 사용되며, 국부적 수용영역에 커널을 곱한 계산 결과는 활성화 맵으로 저장된다. 이러한 공유 속성을 CNN의 가중치 공유(weight sharing)라고 한다[14]. 결과적으로, 다른 커널은 또 다른 활성화 맵을 생성하고, 사용되는 커널의 개수는 하이퍼변수를 통하여 조정할 수 있다. CNN은 가중치 공유(weight sharing)와 로컬 연결(local connectivity) 속성을 결합하여 높은 차원의 데이터를 처리할 수 있다. 즉, CNN에서 매우 적은 수의 뉴런만 후속 층에 연결되며, 이러한 국부적 속성은 완전히 연결된 신경망에 비해 네트워크를 희박하게 연결되게(sparse connection) 만들고, 결과적으로 심층(deep) CNN이 학습이 더 잘 되게 한다. 풀링층(pooling layer)은 일반적으로 컨볼루션층과 다음 층 사이의 중간에 사용된다. 풀링층(pooling layer)은 미리 지정된 풀링 방법으로 입력의 차원을 줄이는 것을 목표로 하며, 가능한 한 많은 정보를 보존하면서 더 작은 입력을 생성한다. 또한 풀링층은 네트워크에 공간 불변성(spatial invariance) 제공하고[30], 이는 모델의 일반화(generalization)를 향상하는 데 도움을 준다. 평균 풀링(averaging pooling), 최소풀링(min-pooling) 및 분수 최대 풀링(fractional max-pooling) 및 확률적 풀링(stochastic pooling)과 같은 다양한 유형의 풀링 방법이 제시됐다. 일반적으로 가장 많이 사용되는 풀링 방법은 불변성을 효율적으로 캡처하여 영상을 처리하는 데 탁월한 효과

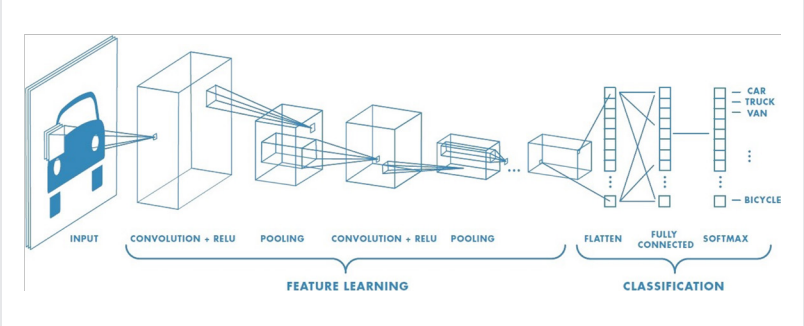

그림 15. 심층 콘볼류션 신경망(Deep convolutional Neural Network, CNN)의 기본적인 구조

를 보이는 최대 풀링(max-pooling) 방법이다[30]. 최대 풀링은 각 지정된 창 내에서 최대값을 추출한다.

현재 CNN은 컴퓨터 비전(computer vision) 작업에서 가장 많이 사용되는 딥러닝 모델이다. CNN은 영상, 비디오 및 음성 데이터의 경우와 같이 데이터 배열의 가까운 값이 서로 상관관계가 있는 배열로 데이터가 구성될 때 매우 효과적인 성능을 발휘한다. 영상 분석에서, CNN의 컨볼루션층(convolutional laryer)은 로컬 연결(local connectivity)과 가중치 공유(weight sharing) 속성을 바탕으로 고차원의 입력을 쉽게 처리할 수 있고, 풀링층(pooling layer)은 필수적인 정보를 잃지 않고 데이터를 다운 샘플링(down-sample)할 수 있다. 따라서 각 컨볼루션층은 다양한 커널을 사용하여 입력 영상을 보다 추상적인 특징 그룹으로 변환할 수 있다. 결과적으로, CNN은 여러 컨볼루션층을 쌓음으로써 입력에서 필수 패턴을 캡처하는 표현으로 변환하여 정확한 예측이 가능하다. 또 다른 한편으로는 CNN은 자연어 처리(natural language processing)에서, 다른 딥러닝 아키텍처에 비해 매우 우수한 결과를 보여준다[31, 32]. 특히, CNN은 텍스트에서 지엽적인 정보를 추출하고, 구와 단어 사이의 의미 및 구문적 의미를 탐색하는 데 능숙하다. 또한 텍스트 데이터의 자연스러운 구성은 CNN 아키텍처에서 쉽게 처리될 수 있다. 따라서 CNN은 최종적인 예측 결과가 입력 텍스트에서 핵심 정보를 추출하는 것에 크게 의존하는 분류 작업을 수행하는 데 있어서, 매우 강력한 잠재력을 보여준다[33].

3) Deep Belief Network(DBN)

Deep Belief Network(DBN)은 서로 다른 유형의 신경망을 결합하여 새로운 신경망 모델을 구성하는 모델 중 하나이다. DBN은 RBM(Restricted Boltzmann Machines)과 D-FFNN(Deep Feedforward Neural Networks)을 통합한 것이다. RBM은 입력 장치를 구성하고 D-FFNN은 출력 장치를 구성한다. RBM은 여러 개를 쌓아서 사용되며, 이는 하나 이상의 RBM이 순차적으로 사용된다는 것을 의미한다. RBM과 DFFNN의 서로 다른 특성으로 인해 두 가지 유형의 학습 알고리즘이 사용된다. 실제로 RBM은 비지도 학습방법으로 모델을 초기화하는 데 사용되고, 매개변수의 미세 조정을 위해 지도학습 방법이 사용된다[34]. 즉, 신경망의 매개변수를 초기화한 후 지도 학습방법으로 미세 조정한다. 사전 훈련 단계에서 생략된 샘플의 레이블이 이때 활용된다. DBN은 자연어 처리[35], 음향 모델링[36], 영상 인식[20] 및 컴퓨터 생물학[37]과 같은 많은 응용 분야에서 성공적으로 사용되고 있다.

4) Autoencoder

Autoencoder는 특징 추출 또는 차원 축소와 같은 표현 학습에 사용되는 비지도 학습 신경망 모델이다. Autoencode의 일반적인 특징은 입력 및 출력 층의 크기가 같으며 대칭적인 구조이다[20]. 기본 아이디어는 입력 패턴 x에서 새로운 인코딩 c = h(x)로의 매핑을 학습하는 것이다. 이는 입력 패턴과 동일한 출력 패턴을 재현하는 것이다. 따라서 일반적으로 x보다 차원이 낮은 인코딩 c를 이용하여 x를 재현할 수 있다. Autoencoder의 구성은 DBN과 유사하다. Autoencoder의 원래 구현[20]은 RBM으로 네트워크의 전반부만 사전 훈련시킨 다음, 네트워크를 확장하여 네트워크의 두 번째 부분을 생성한다. DBN과 유사하게 사전 훈련 단계 다음에는 미세 조정 단계가 뒤따른다. Autoencoder는 레이블을 사용하지 않으므로 비지도 학습 모델이다. Autoencoder는 여러 응용 프로그램에서 차원 축소에 성공적으로 사용되고 있다. Autoencoder는 적절한 양의 데이터를 사용할 수 있을 때 데이터의 훨씬 더 나은 표현을 달성할 수 있다[20]. 차원 축소에서 PCA

가 선형변환인 반면 Autoencoder는 비선형변환이다. 나중에 딥러닝에서 차원 축소 데이터를 사용하면 성능이 향상된다. 이후에, 희소 Autoencoder(sparse autoencoder), 잡음 제거 Autoencoder(denoising autoencoder) 또는 변형 Autoencoder(variational autoencoder)와 같은 다양한 확장 모델이 발표됐다[38, 39, 40].

5) Long Short-Term Memory(LSTM) Network

Long short-term memory network(LSTM)는 1997년 Hochreiter와 Schmidhuber에 의해 처음 소개됐다[19]. LSTM은 데이터의 장기적 종속성을 처리할 때, 잘 수행되지 않는 RNN의 단점을 해결할 수 있는 RNN의 변형 모델이다. DBN과 CNN이 순방향 신경망인 반면에, RNN의 연결에서는 현재의 출력이 다음 학습에서 재사용(피드백 연결)되며, 이를 통해 시간 경과에 따른 동적변화를 모델링할 수 있다[15]. 또한 LSTM은 기울기 소실(gradient vanishing) 또는 기울기 폭발(gradient exploding) 문제를 방지할 수 있다[19]. 1999년에는 셀 메모리를 재설정할 수 있는 망각 게이트(forget gate)가 있는 LSTM이 도입됐다. 이것은 초기 LSTM을 개선하고 LSTM 네트워크의 표준 구조가 됐다[41]. LSTM은 벡터나 배열과 같은 단일 데이터 포인트 뿐만 아니라, 데이터 시퀀스도 처리할 수 있다. 이러한 이유로 LSTM은 음성 또는 비디오 데이터를 분석하는 데 특히 탁월한 성능을 보인다.

3. 딥러닝(deep learning)의 미래 방향

데이터 과학의 모든 모델은 추론 모델(inferential model) 또는 예측 모델(prediction model)로 분류될 수 있다. 추론 모델(inferential model)은 예측을 할 뿐만 아니라 해석 가능한 구조를 제공하며, 이는 예측 프로세스 자체의 모델이 된다. 인과 모델(causal model)이 여기에 해당한다. 이에 비해 예측 모델(prediction model)은 예측을 위한 블랙박스 모델(Black-Box Model)이다. 최근에 해석 가능한 또는 설명 가능한 AI(interpretable or explainable AI, XAI)에 대한 관심이 높아지고 있다(그림 16)[42, 43]. 특히 임상 및 의료 분야에서는 환자의 설득력을 높이기 위해서 통계적 예측에서 이해 가능한 모델이 필요하다[44]. 현재는, XAI 모델과 설명할 수 없는 모델의 구분이 잘 정의되어 있지 않다. XAI 분야는 아직 초기 단계이지만 일반적인 딥러닝 모델에 대한 의미 있는 해석이 가능할 수 있다면, 이는 이 분야에 혁신을 일으킬 수 있다.

일반적으로, 정확한 통계분석을 위해서, 실험 설계 단계에서 사용 가능한 표본 크기가 특정 통계분석을 수행하기에 충분한지 평가한다. 반면에, 딥러닝 방법에서는 충분한 표본을 의미하는 빅 데이터를 이용하여 학습하고 있다고 가정한다. 이는 이상적인 경우에 해당하며, 실제 딥러닝 응용의 경우, 사용 가능한 데이터 표본 크기가 딥러닝 모델을 학습하기에 충분한지 사례별로 확인해야 한다. 일반적으로 딥러닝 모델은 수만 개 이상의 표본이 확보되면 좋은 성능을 보이지만, 적은 수의 데이터에서 잘 수행되는지는 대체로 불분명하다. 이 문제를 입증하기 위한 예로서, EMNIST 데이터의 분류 정확도에 대한 표본 크기의 영향을 조사하기 위해 연구가 수행됐다[45]. EMNIST(Extended MNIST)는 0-9의 필기체 숫자 10개의 클래스에 대해 280,000개의 필기체 문자(240,000개의 훈련 표본 및 40,000개의 테스트 표본)로 구성된다. 10개 클래스의 필기체 숫자 분류 작업에 다층 LSTM 모델을 사용하여 분석하였고, 분석 결과에서 5% 미만의 분류 오류를 달성하려면 25,000개 이상의 훈련 표본이 필요함이 보고됐다[45]. 이러한 결과는 딥러닝 모델이 작은 데이터에 대하여 기적을 일으킬 수 없음을 보여준다. 따라서, 표본 크기

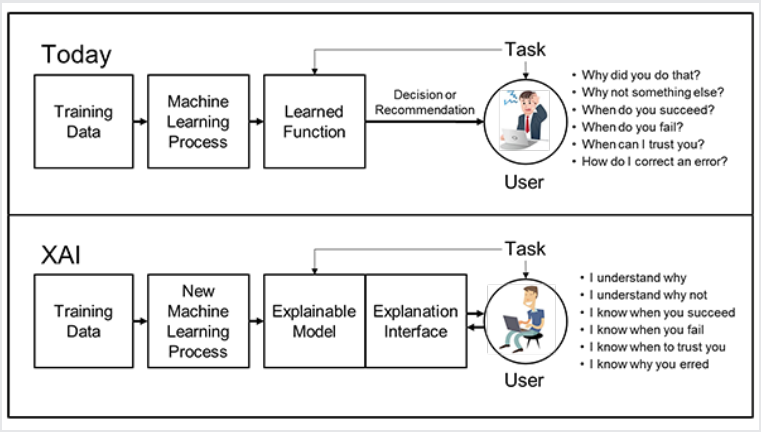

그림 16. 해석 가능한 또는 설명 가능한 AI(XAI) 개념

가 너무 작으면 딥러닝이 제대로 작동하지 못하며, 특정 작업에 딥러닝을 사용할 때, 적합한 딥러닝 모델과 데이터의 조합은 매우 중요하다.

위에서 제시된 핵심 아키텍처 외에도, 추가적인 딥러닝 네트워크의 발전된 모델이 존재한다. 예를 들어, 딥러닝과 강화학습(reinforcement learning)은 서로 결합되어 심층 강화학습을 형성한다[46,47,48]. 이러한 모델은 로봇, 게임 및 의료 분야에서 응용되고 있다. 고급 모델의 또 다른 예로는 데이터가 그래프 형식일 때 특히 적합한 그래프 CNN(graph CNN)이 있다[49,50]. 이러한 모델은 자연어 처리, 추천 시스템, 유전체학 및 화학 분야에서 사용되고 있다. 또 다른 발전된 모델은 VAE(Variational Autoencoder)이다[51,52]. VAE는 입력에 대한 인코딩으로 잠재 공간에 대한 분포를 사용하는 조정된 Autoencoder이다. VAE는 영상 또는 텍스트 생성을 위해서 비지도 학습방식으로 유사한 데이터를 생성하기 위해 사용된다. 애플리케이션에 적합한 딥러닝 모델을 찾는 데 있어서의 문제는 모델이 응용되는 분야가 서로 배타적이지 않다는 것이다. 즉 딥러닝 모델 간 상당한 중복이 있으며, 대부분의 경우 비교연구를 수행해야만 최상의 모델을 찾을 수 있다[53].

결론적으로, 딥러닝 모델의 신경망 아키텍처는 레고와 같은 구성을 허용하는 유연성을 제공한다. 즉 여기서 논의된 핵심 아키텍처 빌딩 블록의 요소를 활용하여 딥러닝 모델을 무제한으로 구성할 수 있다. 따라서 향후 딥러닝의 발전적 응용을 위해서는 이러한 핵심 요소에 대한 기본적 이해는 반드시 필요하다. 향후, 미래에 향상된, 투명한, 안전한, 공정하고 편향되지 않은 AI 학습 모델을 얻기 위해서는, 설명 가능한 AI (XAI)가 없어서는 안될 실용적인 요소가 될 것이다.

참 고 문 헌

1. McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115-133.
2. Hebb, D. (1949). The Organization of Behavior. New York, NY:Wiley.
3. Rosenblatt, F. (1957). The Perceptron, A Perceiving and Recognizing Automaton Project Para. Cornell Aeronautical Laboratory.
4. Widrow, B., and Hoff,M. E. (1960). Adaptive Switching Circuits. Technical Report, Stanford University, California; Stanford Electronics Labs.
5. Ivakhnenko, A. G. (1968). The group method of data of handling; a rival of the method of stochastic approximation. Soviet Autom. Control 13, 43-55.
6. Ivakhnenko, A.G. (1971). Polynomial theory of complex systems. IEEE Trans. Syst. Man Cybernet. SMC-1, 364-378.
7. Minsky, M., and Papert, S. (1969). Perceptrons. MIT Press.
8. Werbos, P. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences (Ph.D. thesis), Harvard University, Harvard, MA, United States.
9. Werbos, P. J. (1981). "Applications of advances in nonlinear sensitivity analysis," in Proceedings of the 10th IFIP Conference, 31.8-4.9, New York, 762-770.
10. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernet. 36, 193-202.
11. Fukushima, K. (2013). Training multi-layered neural network neocognitron. Neural Netw. 40, 18-31. doi: 10.1016/j.neunet.2013.01.001.
12. Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. 79, 2554-2558.
13. Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. Nature 323, 533-536.
14. Le Cun, Y. (1989). Generalization and Network Design Strategies. Technical Report CRG-TR-89-4, Connectionism in Perspective. University of Toronto Connectionist Research Group, Toronto, ON.
15. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521:436.
16. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. Neural Comput. 1, 541-551.
17. Hochreiter, S. (1991). Untersuchungen zu Dynamischen Neuronalen Netzen. Diploma, Technische Universität München 91.
18. Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. Neural Comput. 4, 234-242.
19. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9, 1735-1780.
20. Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science 313, 504-507. doi: 10.1126/science. 1127647.
21. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). ImageNet Classification with Deep Convolutional Neural Networks. Curran Associates, Inc.
22. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b). "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 1097-1105.
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in Advances in Neural Information Processing Systems, 2672-2680.
24. Duda, R. O., Hart, P. E., and Stork, D. G. (2000). Pattern Classification. 2nd Edn. Wiley.
25. Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural Netw. 4, 251-257.
26. Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). "The expressive power of neural networks: a view from the width," in Advances in Neural Information Processing Systems, 6231-6239.
27. Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent," in Proceedings of COMPSTAT'2010 (Springer), 177-186.
28. Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). Deeptox: toxicity prediction using deep learning. Front. Environ. Sci. 3:80. doi: 10.3389/fenvs.2015.00080.
29. Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on chembl. Chem. Sci. 9, 5441-5451. doi: 10.1039/C8SC00148K.
30. Scherer, D.,Müller, A., and Behnke, S. (2010). "Evaluation of pooling operations in convolutional architectures for object recognition," in International Conference on Artificial Neural Networks (Springer), 92-101.
31. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv [Preprint]. arXiv:1408.5882. doi: 10.3115/v1/D14-1181.
32. Yang, Z., Dehmer, M., Yli-Harja, O., and Emmert-Streib, F. (2020). Combining deep learning with token selection for patient phenotyping from electronic health records. Sci. Rep. 10:1432. doi: 10.1038/s41598-020-58178-1.
33. Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. arXiv [Preprint]. arXiv:1702.01923.
34. Yoshua, B. (2009). Learning deep architectures for AI. Foundat. Trends Mach. Learn. 2, 1-127. doi: 10.1561/2200000006.
35. Sarikaya, R., Hinton, G. E., and Deoras, A. (2014). Application of deep belief networks for natural language understanding. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 778-784. doi:

참 고 문 헌

10.1109/TASLP.2014.2303296.
36. Mohamed, A.-R., Dahl, G. E., and Hinton, G. (2011). Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20, 14-22. doi: 10.1109/TASL.2011.2109382.
37. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., et al. (2015). A deep learning framework for modeling structural features of rna-binding protein targets. Nucleic Acids Res. 43:e32. doi: 10.1093/nar/gkv1025.
38. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. -A. (2010). Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371-3408. Available online at: http://www.jmlr.org/papers/v11/vincent10a.html.
39. Deng, J., Zhang, Z.,Marchi, E., and Schuller, B. (2013). "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (IEEE), 511-516.
40. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., et al. (2016). "Variational autoencoder for deep learning of images, labels and captions," in Advances in Neural Information Processing Systems, 2352-2360.
41. Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: continual prediction with LSTM. Neural Comput. 12, 2451-2471. doi: 10.1162/089976600300015015.
42. Biran, O., and Cotton, C. (2017). "Explanation and justification in machine learning: a survey," in IJCAI-17Workshop on Explainable AI (XAI). Vol. 8, 1.
43. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv [Preprint]. arXiv:1702.08608.
44. Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv [Preprint]. arXiv:1712.09923.
45. Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. (2017). Emnist: an extension of mnist to handwritten letters. arXiv[Preprint]. arXiv:1702.05373. doi: 10.1109/IJCNN.2017.7966217.
46. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. Nature 518:529. doi: 10.1038/nature14236.
47. Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: a brief survey. IEEE Signal Process. Mag. 34, 26-38. doi: 10.1109/MSP.2017.2743240.
48. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., andMeger, D. (2018). "Deep reinforcement learning that matters," in Thirty-Second AAAI Conference on Artificial Intelligence.
49. Henaff, M., Bruna, J., and LeCun, Y. (2015). Deep convolutional networks on graph-structured data. arXiv [Preprint]. arXiv:1506.05163.
50. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. arXiv [Preprint]. arXiv:1901.00596.
51. An, J., and Cho, S. (2015). Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability. Special Lecture on IE 2.
52. Doersch, C. (2016). Tutorial on variational autoencoders. arXiv [Preprint]. arXiv:1606.05908.
53. Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. Front Artif Intell. Feb 28;3:4. doi: 10.3389/frai.2020.00004.