홍익대학교

합성곱 신경망(CNN) 및 다양한 딥러닝 모델의 확장

Hanjin Cho

Electronic & Electrical Convergence Engineering
Hongik University
Republic of Korea

강의 내용

HONGIK UNIVERSITY

- 합성곱 신경망(CNN)의 개요
- 합성곱 신경망(CNN)의 구성
- 주요 합성곱 신경망
- 컴퓨터 비전에서의 CNN 응용 분야
- 다양한 딥러닝 모델의 확장

합성곱 신경망(CNN)의 개요

HONGIK UNIVERSITY

- **한계와 필요성: MLP에서 CNN으로**
  - MLP는 2차원 이미지를 1차원 벡터로 변환하여 입력하므로 공간 정보를 손실하여 이미지 인식 성능 제한.
  - 반면, 인간의 시각 시스템은 이미지의 국소적인 영역(수용장, receptive field)에서 특징을 추출함.
  - CNN은 이러한 방식에서 영감을 받아, 2차원 구조를 유지한 채 특징을 추출하도록 설계됨.

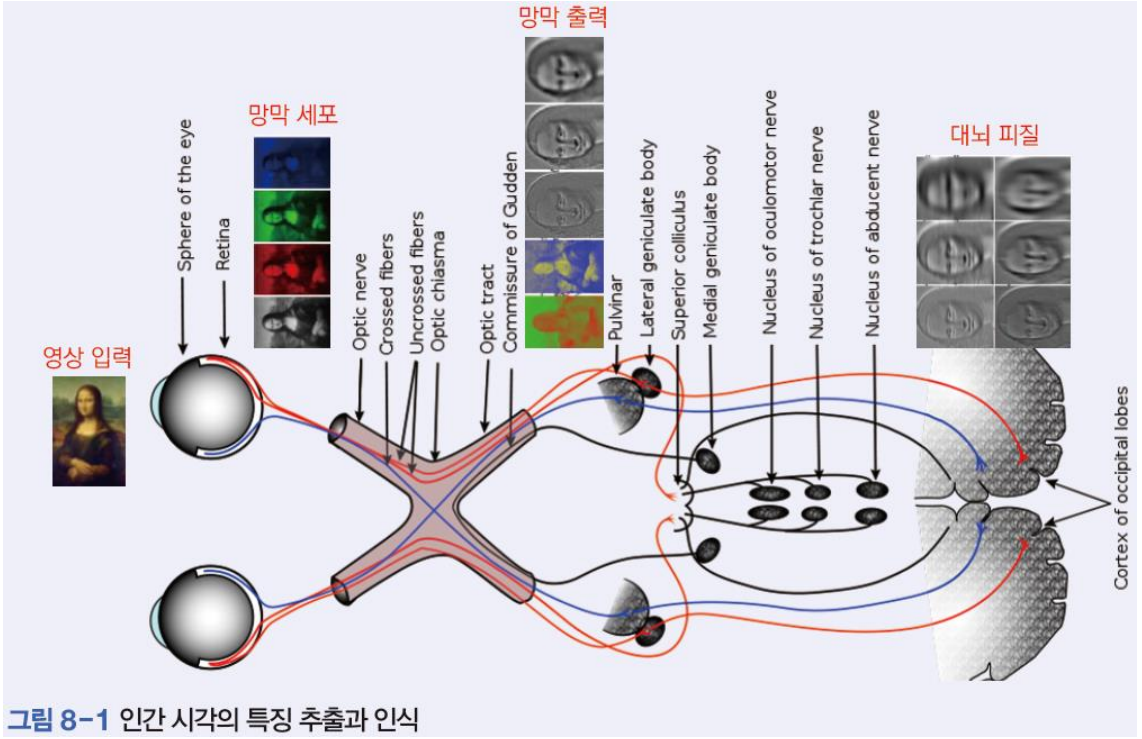
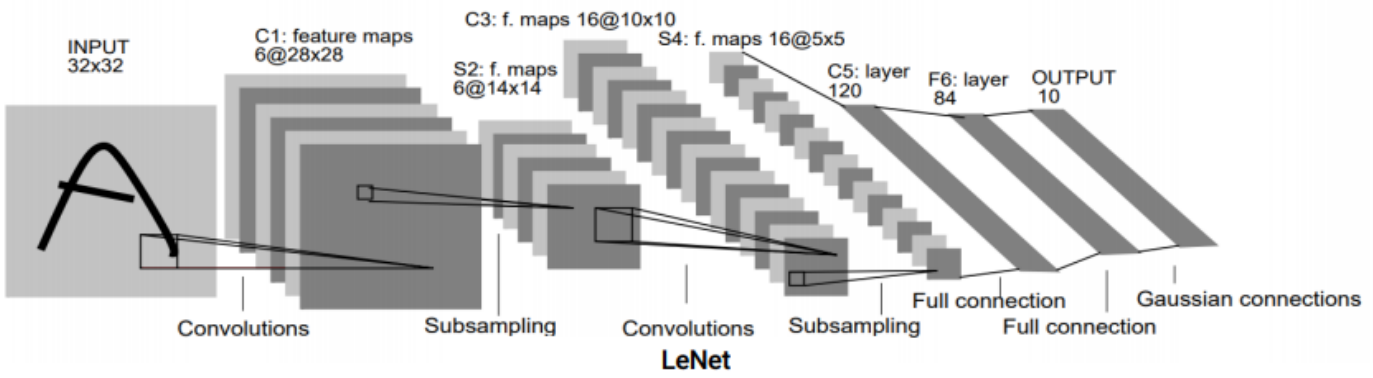


그림 8-1 인간 시각의 특징 추출과 인식

합성곱 신경망(CNN)의 개요

HONGIK UNIVERSITY

- **합성곱 신경망(CNN) 이란?**
  - CNN은 이미지 인식, 영상 분석, 객체 탐지 등 시각적 데이터를 처리하는 데 최적화된 딥러닝 구조임.
  - 데이터의 공간적 구조를 인식하고 학습할 수 있도록 설계됨.
  - 주요 구성
    - 합성곱 층(Convolutional Layer), 풀링 층(Pooling Layer), 완전연결 층(Fully Connected Layer).
  - MLP보다 적은 파라미터 수로 효율적인 학습이 가능함.



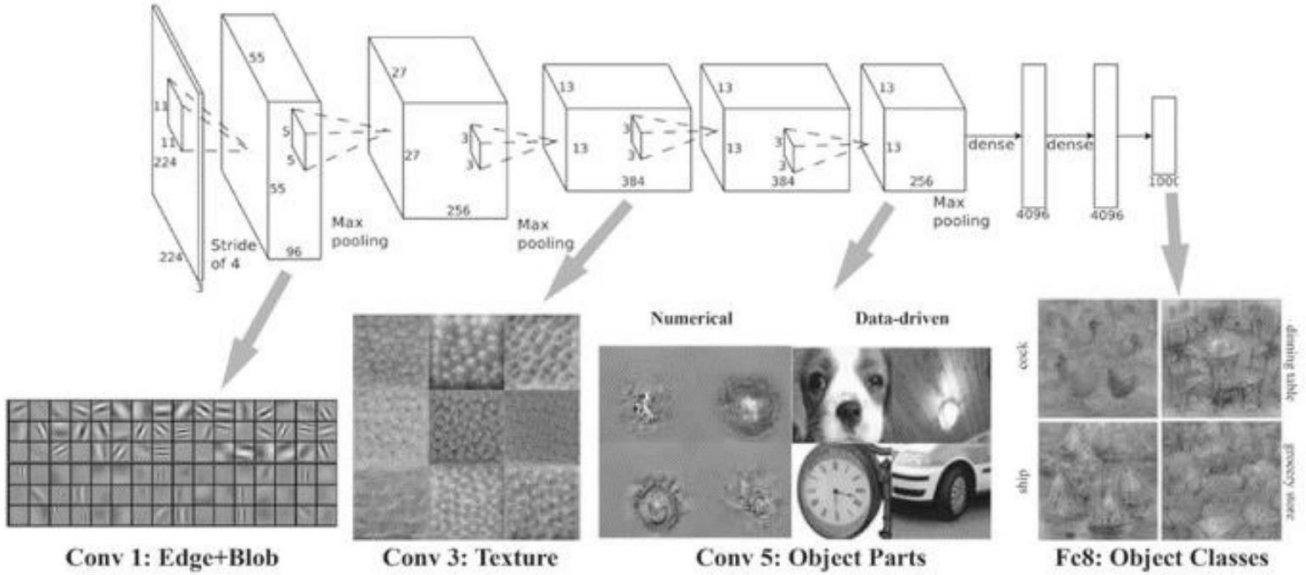
LeNet

LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

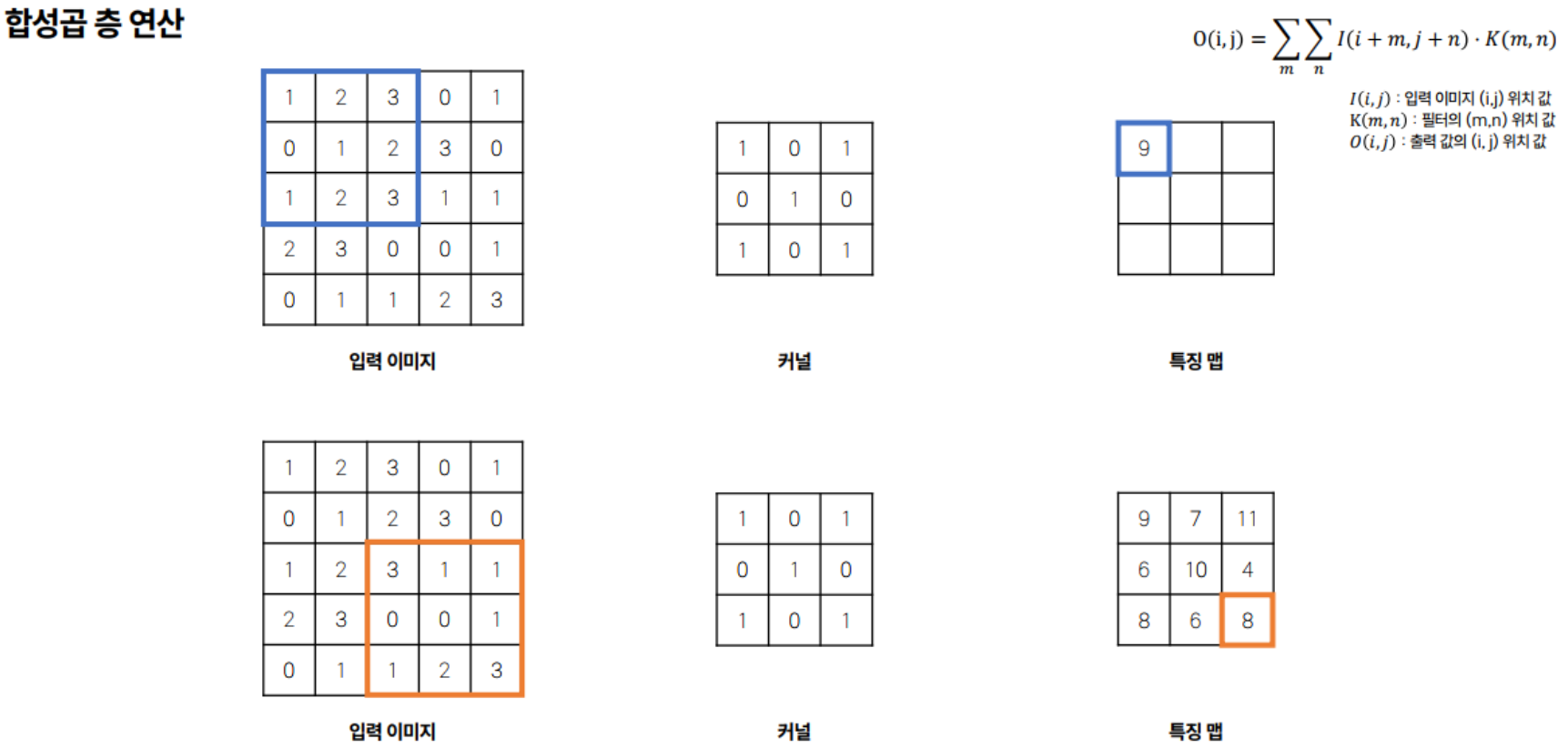
- **합성곱 층 (Convolutional Layer)**
  - 합성곱 층은 이미지의 공간 구조를 유지한 채 유용한 특징을 추출함.
  - 입력 데이터에 필터(커널)를 적용해 합성곱 연산을 수행하고, 특징맵(feature map)을 생성함.
  - 작은 영역을 슬라이딩하며 곱셈-합산을 수행해 국소적인 패턴을 감지함.
  - 자주 사용하는 커널 크기: 1×1, 3×3, 5×5, 7×7 등.



합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

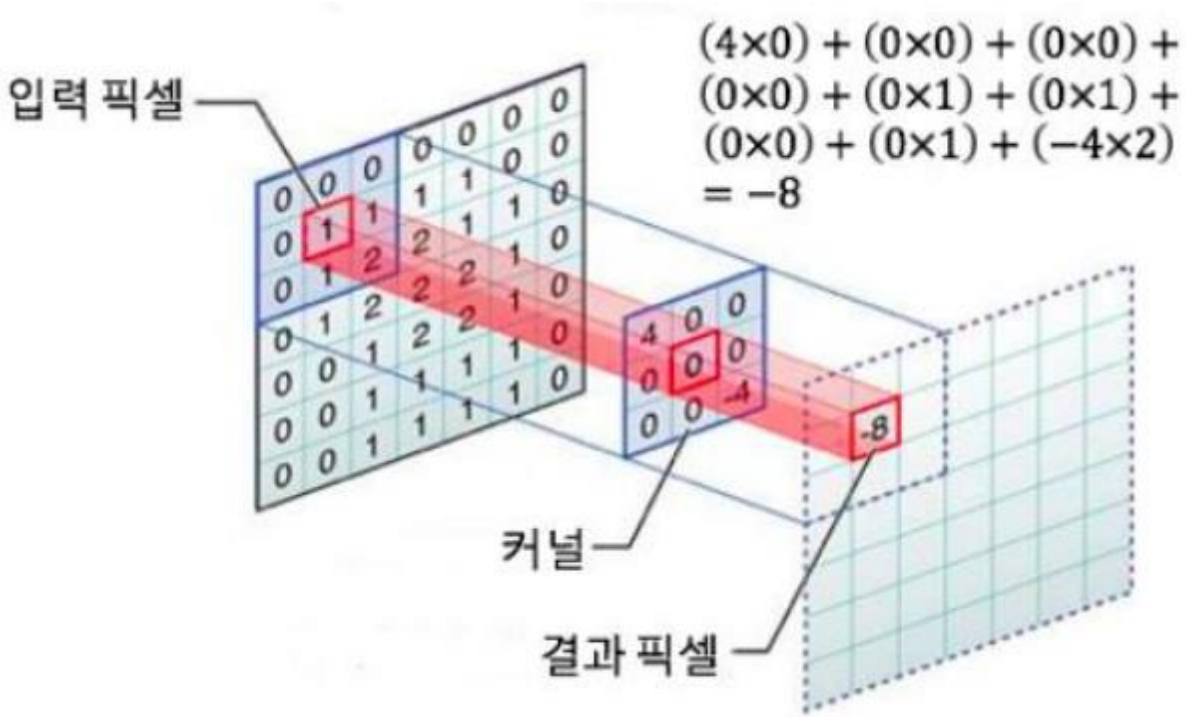
- **합성곱 연산 예시**
  - 입력 이미지의 국소 영역과 커널 간 곱셈 후 합산하여 특징 맵의 값을 계산함.
  - 입력 이미지가 커널과 겹치는 부분마다 이 연산 반복하여 전체 특징 맵 생성.



합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

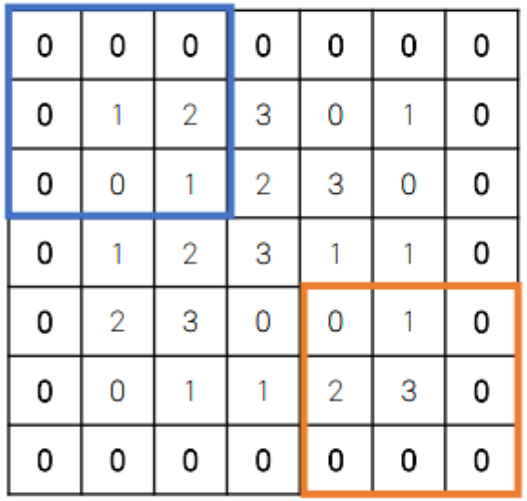
- **합성곱 연산 시각화**
  - 입력 픽셀과 커널 간 곱셈 결과를 통해 특징을 추출하는 구조 시각화.
  - 각 커널은 선, 에지, 방향 등의 저수준 특징을 추출하도록 설계됨.
  - 여러 개의 커널이 동시에 다양한 특징을 병렬적으로 추출함.



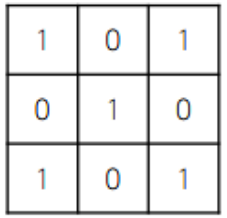
합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

- **패딩의 역할**
  - 합성곱 연산 후 출력 크기가 줄어드는 것을 방지하거나 유지하기 위해 패딩을 사용함.
  - 일반적으로 0으로 채우는 제로 패딩 사용.
  - 입력 가장자리에 추가적인 픽셀을 더해 출력 크기를 유지하거나 특정 크기를 맞춤.



입력 이미지 (제로패딩 적용)



커널

특징 맵

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

- **스트라이드(합성곱 층 이동 간격)와 다채널 처리**
  - 스트라이드는 커널이 입력 이미지 위를 이동하는 간격을 의미.
  - 기본값은 1이며, 값이 클수록 출력 특징 맵의 크기는 작아짐.
  - 다채널 입력(RGB 등)일 경우, 각 채널마다 개별 필터가 적용되고, 그 결과를 합산해 최종 출력 생성.

$$O(i,j) = \sum_m \sum_n I(i \cdot S + m, j \cdot S + n) \cdot K(m,n)$$

$I(i,j)$: 입력 이미지 (i,j) 위치 값
$K(m,n)$: 필터의 (m,n) 위치 값
$O(i,j)$: 출력 값의 (i, j) 위치 값
S: 스트라이드 값

1	2	3	0	1
0	1	2	3	0
1	2	3	1	1
2	3	0	0	1
0	1	1	2	3

입력 이미지 (스트라이드 2로 설정)

1	0	1
0	1	0
1	0	1

커널

9	11
8	8

특징 맵

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

- **출력 크기(특징 맵 크기) 계산**
  - 필터 크기(F), 패딩(P), 스트라이드(S)를 기반으로 출력 크기 계산:

$$O_w = \frac{W - F + 2P}{S} + 1 = \frac{28 - 3 + 0}{1} + 1 = 26$$

$$O_h = \frac{H - F + 2P}{S} + 1 = \frac{28 - 3 + 0}{1} + 1 = 26$$

W, H : 입력이미지의 너비와 높이
F : 필터(커널)의 크기
P : 패딩 크기
S : 스트라이드

(예) 28×28 이미지에 3×3 필터를 적용, 패딩 없음, 스트라이드 1일 경우 → 26×26 출력.

- 가장 일반적으로 사용되는 필터 크기는 3×3이며, 이는 공간적 국소 패턴 인식에 효과적임.

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

▪ 풀링 층

- 풀링층은 CNN에서 출력 feature map의 공간 크기를 줄이는 역할을 수행함.
- 계산량을 줄이고 과적합을 방지하며, 위치 변화에 대한 모델의 강건함을 높임.
- 중요한 정보를 유지하면서 크기를 축소하고 연산 효율성을 높이기 위해 사용됨.

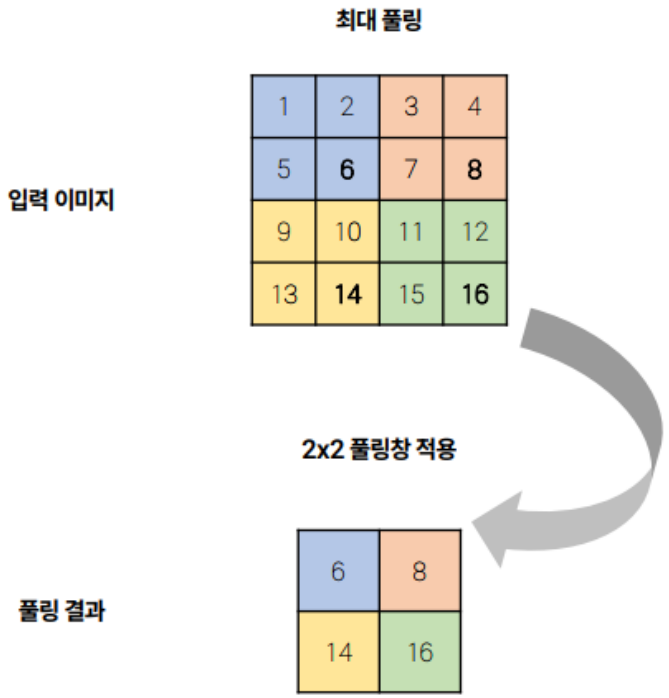
- 풀링 종류
  - 최대 풀링 (Max Pooling)
    → 지정된 창(window) 내에서 가장 큰 값을 선택해 출력.
    → 주요 특징을 강조하고, 작은 변동의 영향을 줄이는 데 유리함.

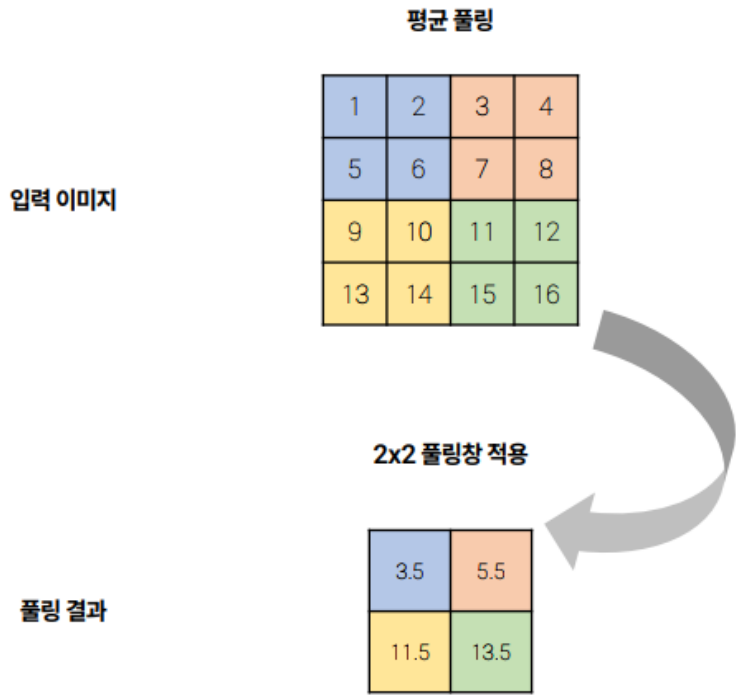
  - 평균 풀링 (Average Pooling)
    → 창 내의 값 평균을 출력.
    → 전체적인 흐름이나 배경 정보를 유지하는 데 유리함.

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

- **풀링 층 예시**
  - 2x2 필터를 사용하여 입력 이미지를 4개의 블록으로 나누고 각각에 대해 풀링 수행.
    - Max Pooling 결과: 각 블록의 최댓값으로 구성.
    - Average Pooling 결과: 각 블록의 평균값으로 구성.
  - 시각적으로 풀링 연산이 어떻게 공간 크기를 줄이는지를 보여줌.





합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

▪ 완전 연결 계층

- CNN의 마지막 단계에서 사용되며, 추출된 특징을 종합하여 최종 예측 수행.
- 모든 입력 뉴런이 모든 출력 뉴런과 연결된 구조를 가짐.
- 합성곱과 풀링 결과는 주로 2D 텐서이므로 완전 연결층 입력을 위해 1D 벡터로 변환(Flattening)이 필요.
- 최종 출력층에서는 이미지의 클래스 혹은 예측값 결정.

(예) 분류 문제 → Softmax 사용
회귀 문제 → 활성화 함수 사용 X

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

▪ 전이학습과 미세조정

- 전이학습은 사전 학습된 모델을 새로운 작업에 활용하는 방식.
- 일반적으로 ImageNet 등 대규모 데이터셋으로 학습된 CNN 모델(VGG, ResNet 등)을 사용함.
- 저수준 특징(엣지, 색상 등)부터 고수준 특징(형태, 패턴 등)을 포함하여 새로운 문제에도 효과적으로 적용 가능함.
- 활용 방식
  - 기존 모델의 가중치를 freeze하고 마지막 분류 계층만 재학습.
  - 또는 전체 계층 중 일부를 fine-tuning하여 새로운 데이터셋에 맞게 조정.
  - 데이터셋이 원래와 크게 다르지 않으면 전이학습만으로도 높은 성능 확보 가능.

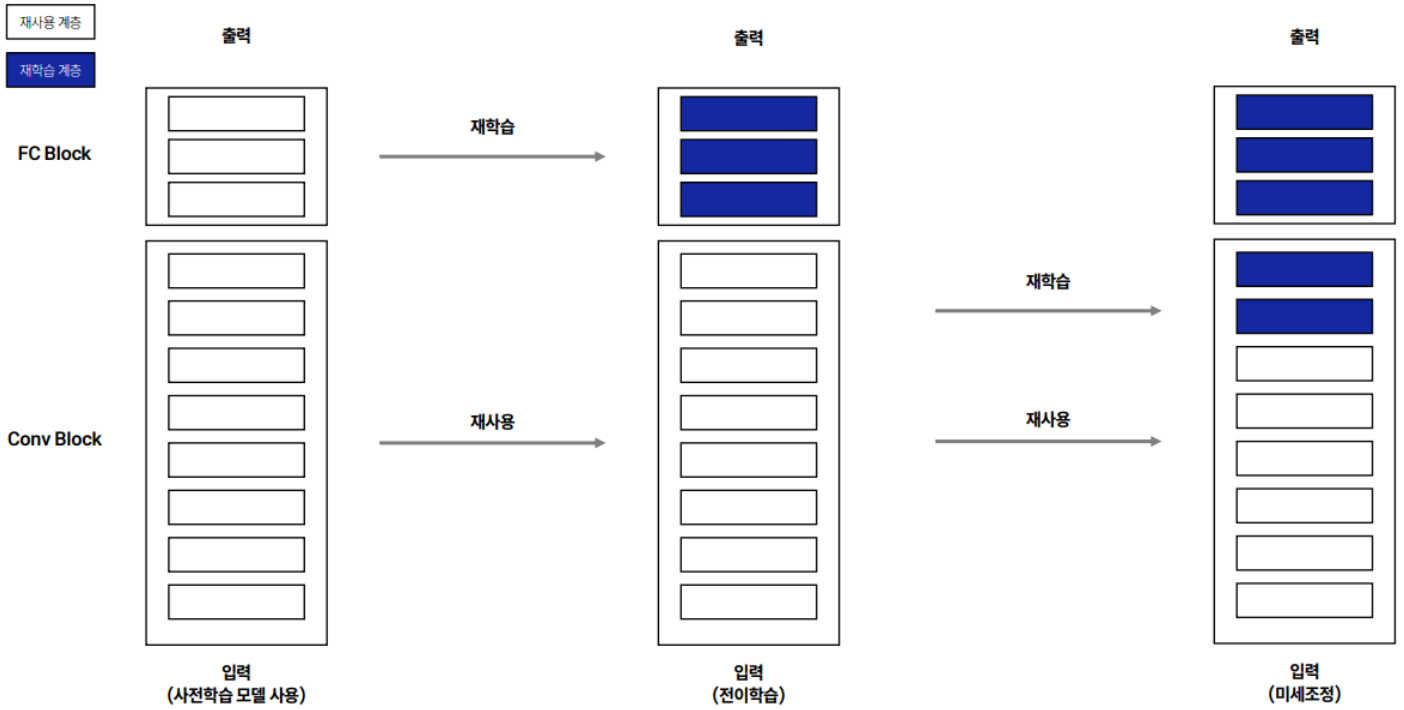
※ 실습 안내

- CNN 기본 코드를 기반으로 전이학습을 적용한 실습을 진행함.
- 사전학습 모델을 활용하여 전처리 및 분류 성능 튜닝 과정을 직접 수행할 예정.

합성곱 신경망(CNN)의 구성

HONGIK UNIVERSITY

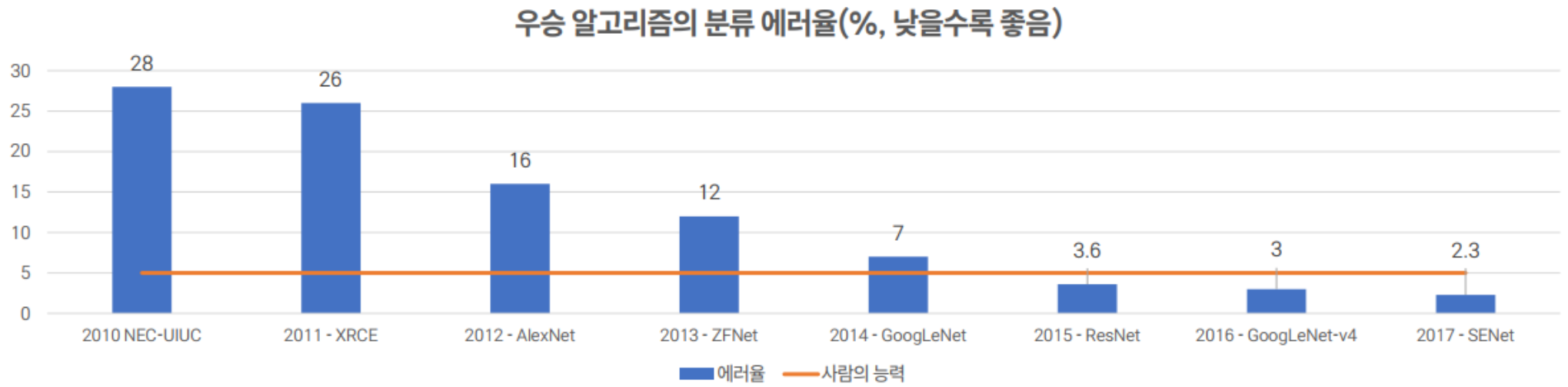
- **전이학습 구조도**
  - Conv Block은 사전 학습된 모델을 그대로 재사용함.
  - [전이학습] : FC Block은 분류 계층으로, 이 부분만 재학습함.
  - [미세조정(Fine-Tuning)] : 일부 Conv Block까지 재학습할 수 있음.
  - 계층 구성에 따라 전이학습 → 미세조정 순서로 성능 개선이 가능함.



주요 합성곱 신경망

HONGIK UNIVERSITY

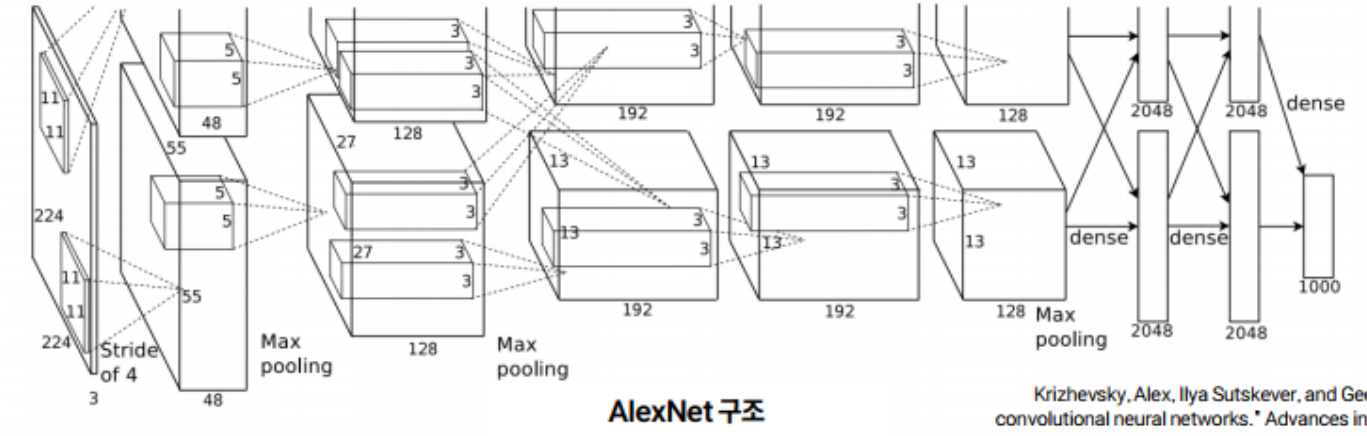
- **ILSVRC로 인해 이미지 인식 성능 대폭 향상**
  - ILSVRC(ImageNet Large Scale Visual Recognition Challenge)는 이미지 분류와 객체 인식 성능을 평가하는 대표 대회.
  - ImageNet은 수백만 장의 이미지와 수천 개의 클래스로 구성됨.
  - 2010~2017년까지 매년 개최되었으며, 이후 분류 성능이 인간 수준(오류율 5%) 이하로 도달하면서 중단.
  - 2017년 SENet의 분류 오류율은 2.3%로 인간보다 낮은 수준을 기록함.



주요 합성곱 신경망

HONGIK UNIVERSITY

- **AlexNet (2012) - CNN의 대중화를 이끈 모델**
  - ILSVRC 2012에서 AlexNet이 우승하며 CNN 기반 접근이 주류로 자리잡음.
  - 총 8개의 층으로 구성되며, 이 중 5개는 합성곱 층, 3개는 완전 연결층으로 구성됨.
  - 모든 층에서 ReLU 활성화 함수를 사용하여 계산 효율성 및 학습 속도 향상.
  - ReLU는 음수를 0으로 변환, 양수는 그대로 유지함.
  - 드롭아웃 기법으로 과적합 방지 및 일반화 성능 향상.
  - GPU 사용을 통해 대규모 데이터셋을 빠르게 학습하여 학습 시간 단축.

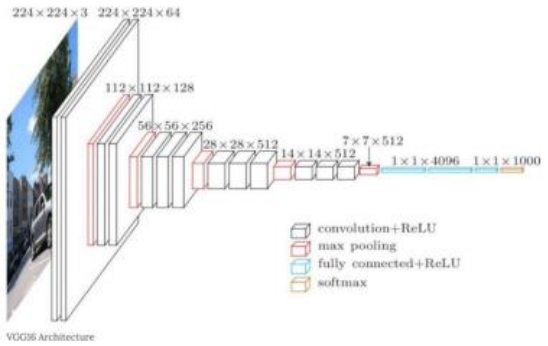


Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

주요 합성곱 신경망

HONGIK UNIVERSITY

- **VGGNet (2014) - 깊은 구조와 단순한 필터로 성능 향상**
  - VGGNet은 Oxford VGG 팀에서 제안한 모델로, VGG16, VGG19 등 다양한 구조로 제공됨.
  - 3x3 필터의 합성곱 층과 2x2 풀링층을 반복적으로 쌓아 단순한 구조로 깊이를 증가시킴.
  - 깊은 네트워크를 통해 고수준의 특징을 추출하고 복잡한 패턴을 효과적으로 학습함.
  - 구조의 단순성과 반복성이 특징으로, 이후 CNN 설계에 널리 영향을 미침.



VGG16 모델 구조

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

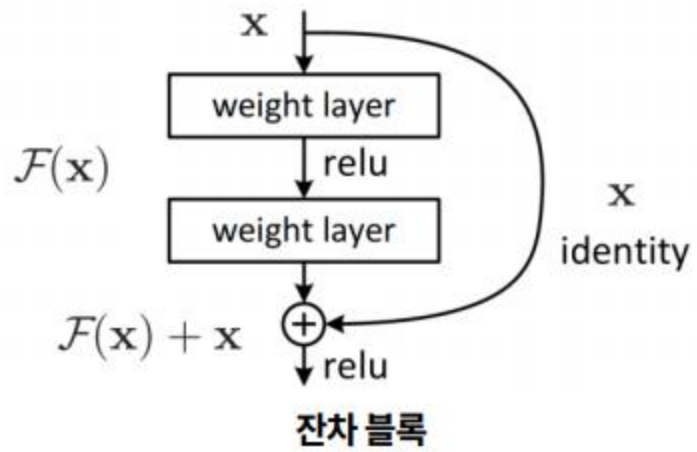
VGG 모델 구성

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

주요 합성곱 신경망

HONGIK UNIVERSITY

- **ResNet (2015) - 잔차 학습을 통한 깊은 네트워크 설계**
  - Microsoft Research에서 제안한 Residual Network.
  - 잔차 학습(residual learning)을 도입하여 깊은 네트워크의 학습 안정성 확보.
  - 핵심 구조인 잔차 블록은 입력값과 출력값을 더하는 skip connection을 포함함.
  - Output = F(x) + x 형태로 구성되어 정보 손실 없이 학습 가능.
  - 깊이에 따른 성능 저하 문제(Degradation Problem)를 효과적으로 해결.
  - ResNet은 50, 101, 152 레이어까지 깊어져도 학습이 가능함을 입증함.



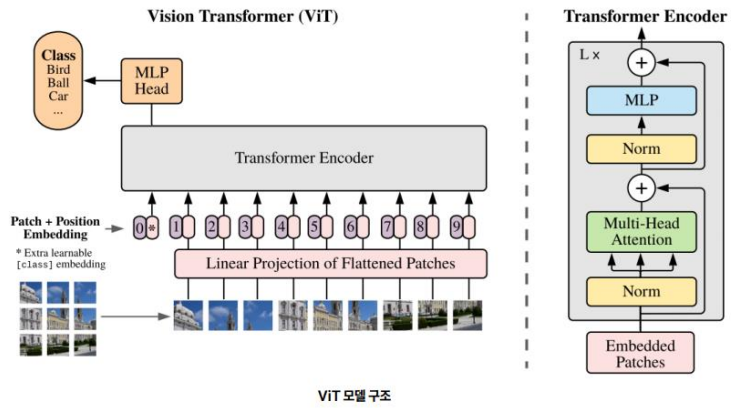
잔차 블록

주요 합성곱 신경망

HONGIK UNIVERSITY

- **Vision Transformers (ViT, 2020) - CNN을 대체하는 새로운 접근**
  - Google에서 제안한 트랜스포머 기반 이미지 분류 모델.
  - 이미지를 고정 크기 패치(예: 16x16)로 나눈 후, 각 패치를 1D 벡터로 변환함.
  - 트랜스포머 인코더가 패치 간 관계를 학습하며, 셀프 어텐션 메커니즘을 활용함.
  - 위치 정보를 제공하기 위해 포지셔널 인코딩을 추가함.
  - CNN은 국소 정보를 중심으로 학습하는 반면, ViT는 전역적인 관계 학습이 가능함.
  - CNN의 한계를 보완하며 새로운 이미지 처리 패러다임을 제시함.

Vision Transformers (2020)



ViT 모델 구조

Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

주요 합성곱 신경망

HONGIK UNIVERSITY

주요 합성곱신경망 요약

모델	AlexNet	VGG	ResNet	ViT (Vision Transformer)
출시연도	2012	2014	2015	2020
주요 아키텍처	5 Conv + 3 FC	16-19 Conv + 3 FC	Residual 블록 (skip connection)	Transformer 블록
파라미터 수	약 60M	약 138M	약 25M (ResNet50)	약 86M
주요 특징	ReLU, Dropout 사용, ImageNet 챌린지에서 우승	작은 3x3 Conv 필터, 매우 깊은 네트워크	깊은 네트워크에서도 성능 저하를 막기 위한 skip connection	이미지 패치를 입력으로 받아 Transformer를 통해 처리
장점	최초의 딥러닝 기반 이미지 분류 네트워크, 간단한 구조	간단한 구조로 쉽게 이해 가능, 성능이 우수	매우 깊은 네트워크 가능, 성능 우수, 거의 모든 문제에 사용 가능	Transformer를 시각적으로 적용한 최초의 성공적 사례
단점	파라미터가 많고 계산 비용이 큼	매우 큰 모델 사이즈, 메모리 및 계산 자원이 많이 필요함	파라미터 수가 여전히 많음	대규모 데이터에 의존, 기존 CNN보다 작은 데이터세트에서 성능 저하

컴퓨터 비전에서의 CNN 응용 분야

HONGIK UNIVERSITY

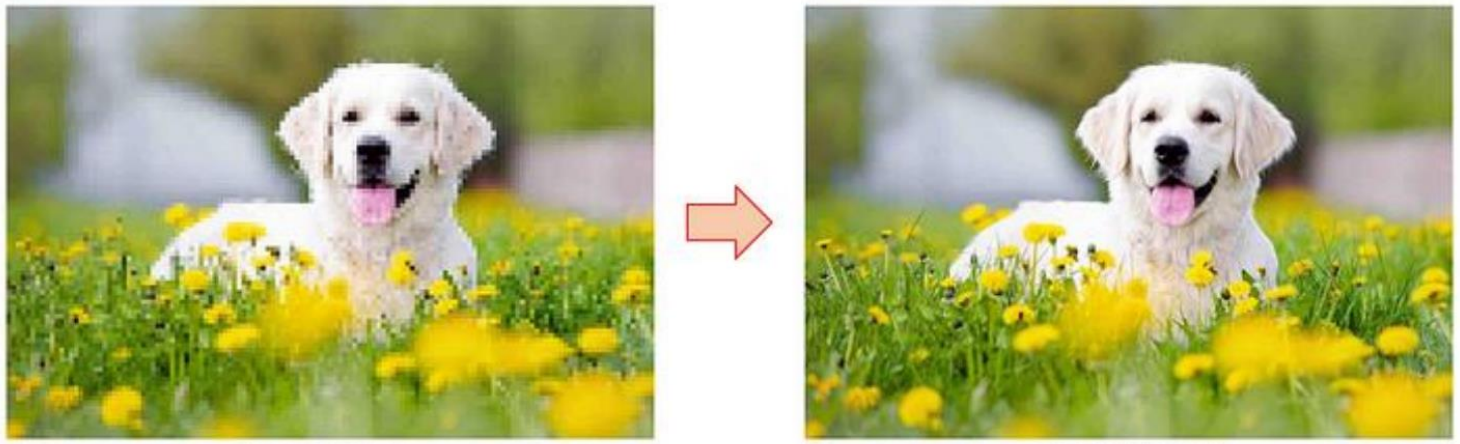
- **주요 비전 과제의 단계별 적용**
  - 분류 (Classification): 이미지 전체를 하나의 클래스로 판단함.
  - 분류 + 위치 추정 (Classification + Localization): 클래스 판단과 함께 위치(박스)를 제시함.
  - 객체 탐지 (Object Detection): 여러 객체를 감지하고, 클래스 및 위치 정보를 함께 제공함.
  - 세그멘테이션 (Segmentation): 픽셀 단위로 객체 영역을 구분함.
  - 분류는 단일 객체, 탐지와 세그멘테이션은 다중 객체 분석에 적합함.

컴퓨터 비전에서의 CNN 응용 분야

HONGIK UNIVERSITY

- **초해상도 (Super Resolution)**
  - 초해상도는 저해상도 이미지를 고해상도로 변환하는 기술.
  - SISR (단일 이미지), MISR (다중 이미지), VSR (비디오 프레임 기반)로 나뉨.
    - SISR: 하나의 이미지만 입력.
    - MISR: 여러 장의 서로 다른 저해상도 이미지를 조합해 복원.
    - VSR: 시간적 연속성을 활용하여 고해상도 프레임을 생성함.

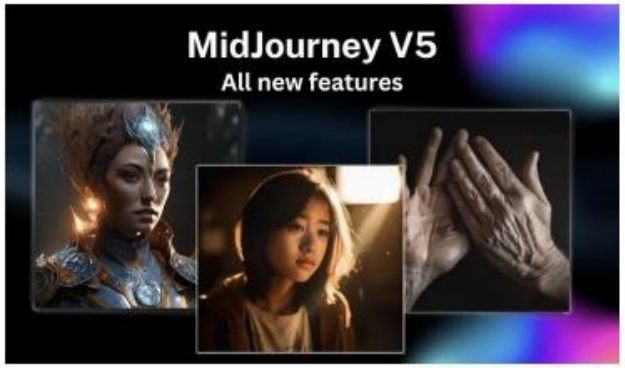
초해상도



다양한 딥러닝 모델의 확장

HONGIK UNIVERSITY

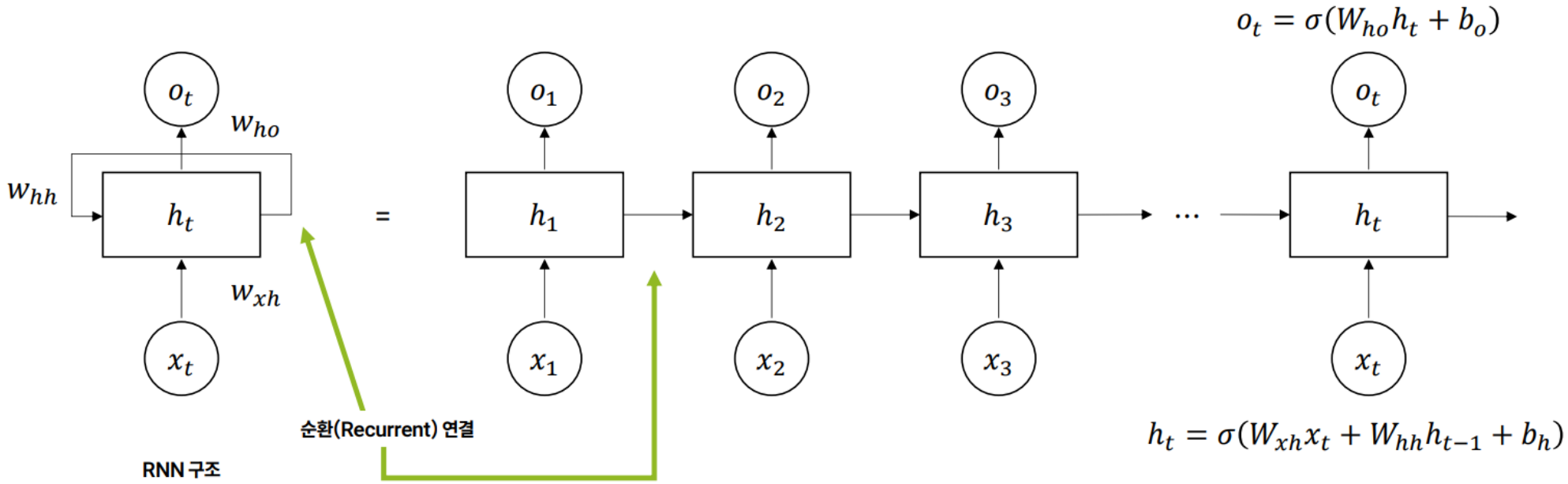
- **이미지 생성형 AI**
  - 텍스트, 그림 등을 기반으로 새로운 이미지를 생성하는 AI 기술.
  - 주로 사용되는 모델:
    - Autoencoder.
    - Variational Autoencoder (VAE).
    - Generative Adversarial Network (GAN).
    - Diffusion 모델.
    - Transformer 기반 모델 등.
  - 다양한 이미지-텍스트 학습을 통해 창의적 이미지 생성 가능.



다양한 딥러닝 모델의 확장

HONGIK UNIVERSITY

- **순환 신경망(Recurrent Neural Network, RNN)**
  - 순환신경망(RNN)은 시간에 따라 변화하는 입력을 처리함.
  - 이전 입력에서 학습한 정보를 다음 단계로 전달하여 순차성을 유지함.
  - 출력뿐만 아니라 은닉 상태가 다음 시점의 입력으로 사용됨.



다양한 딥러닝 모델의 확장

HONGIK UNIVERSITY

- **Transformer**
  - Attention만 사용해 시퀀스를 처리하는 모델.
    - Attention 정보를 전부 다 보는 동시에, 어디에 집중할지를 가중치를 통해 정하는 방법을 의미.
  - RNN 없이도 전체 문맥을 동시에 고려할 수 있음.
  - 포지셔널 인코딩을 통해 단어 순서 정보를 반영함.
  - 인코더는 입력 문장을 의미 있는 벡터로 변환하고, 디코더는 그 벡터를 바탕으로 문장을 생성함.
  - 자연어 처리, 번역, 텍스트 생성 등에 기준 모델로 활용됨.

다양한 딥러닝 모델의 확장

HONGIK UNIVERSITY

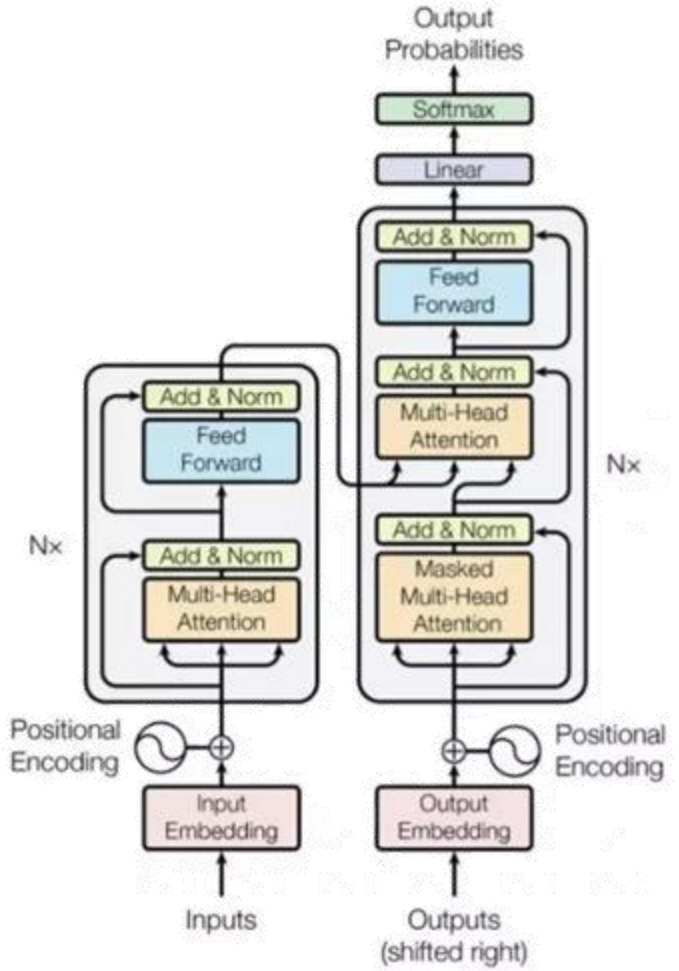
- **텍스트 생성형 AI**
  - 자연어 텍스트 입력을 기반으로 문장을 생성하는 AI.
  - 단어의 확률적 연관성과 문맥을 학습함.
  - 대부분 트랜스포머 기반 구조를 사용.

  - 주로 사용되는 모델:
    - BERT (Bidirectional Encoder Representations from Transformers).
    - GPT (Generative Pre-trained Transformer).
    - HuggingFace Transformers 라이브러리.

다양한 딥러닝 모델의 확장

HONGIK UNIVERSITY

- **대규모 언어모델 (Large Language Model, LLM)**
  - LLM은 수십억 개의 파라미터를 가진 대규모 신경망 언어 모델.
  - 방대한 텍스트 데이터를 기반으로 사전 학습(Pretraining)됨.
  - 질문 응답, 요약, 번역, 창작 등 다양한 자연어 작업 수행 가능.
  - 대표 모델:
    - GPT.
    - Claude.
    - LLaMA.
    - PaLM.
    - Gemini.



  - 현재 생성형 AI의 핵심 기술로,
    대규모 데이터로 학습한 후 다양한 언어 작업을 수행할 수 있는 범용 딥러닝 모델.

Thank you!