TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 12

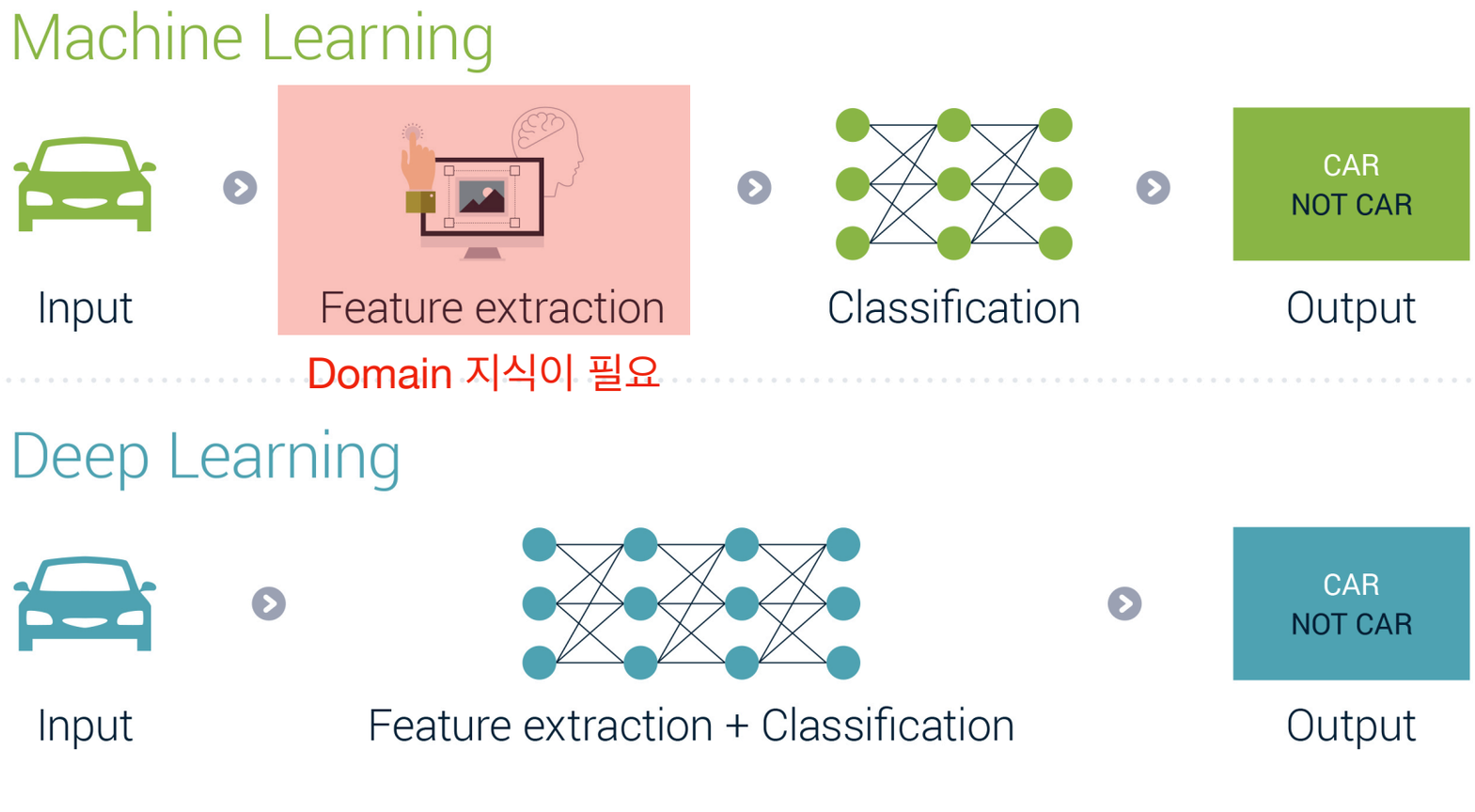
비정형데이터와 머신러닝

연세대학교 | 서중원

Machine Learning

Deep Learning이란?

- 인공 신경망의 확장 버전으로 많은 (깊은) Hidden layer로 이루어진 Deep Neural Network (DNN)을 이용한 학습
- 기존에는 사람이 유의미한 특징을 추출한뒤 학습시켜왔다면, DNN에서는 그 과정이 생략됨.
- 전처리가 (Preprocessing) 생략되는 것은 아님

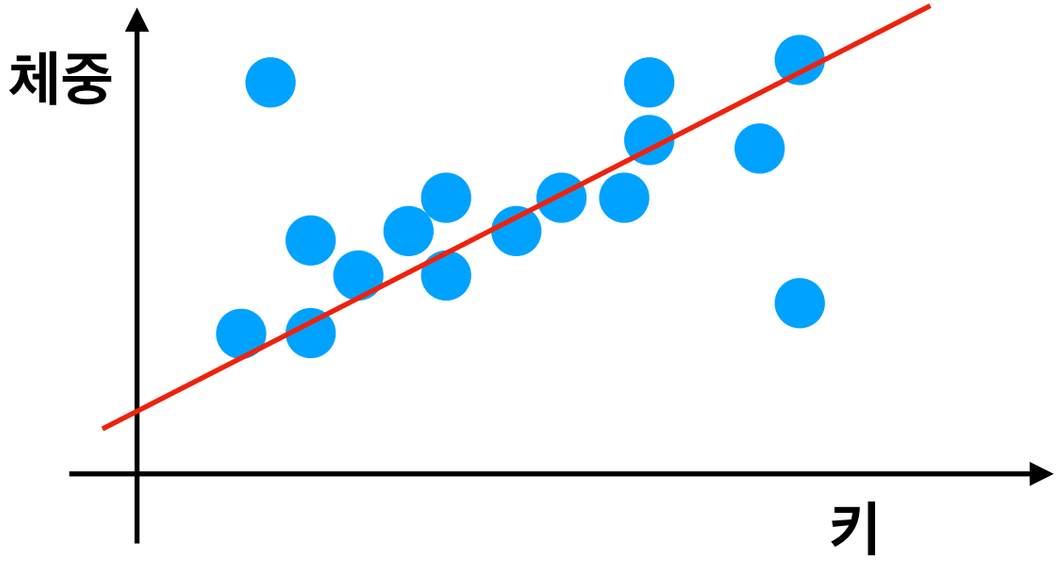


*Source : Jochem Grietens, What's the difference between Machine Learning & Deep (Machine) Learning?, 2018.04.12., https://verhaert.com/difference-machine-learning-deep-learning/.

Machine Learning

선형 회귀 (Linear Regression)

- 선형 회귀는 종속변수 y와 한 개 이상의 독립 변수X와의 선형 상관 관계를 모델링하는 회귀분석 기법이다.

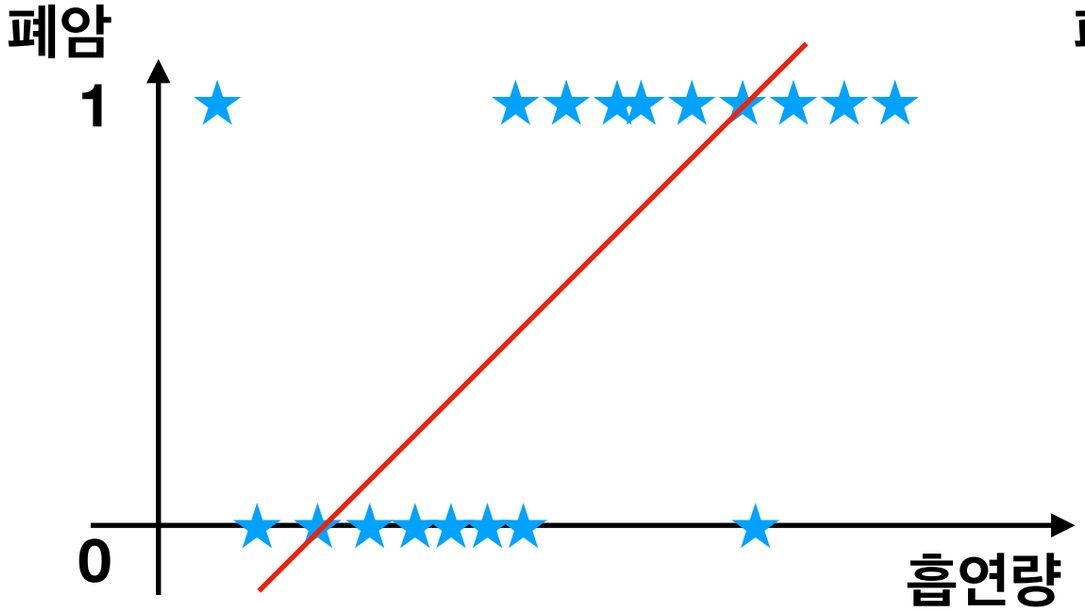


y(체중) = aX(키) + b

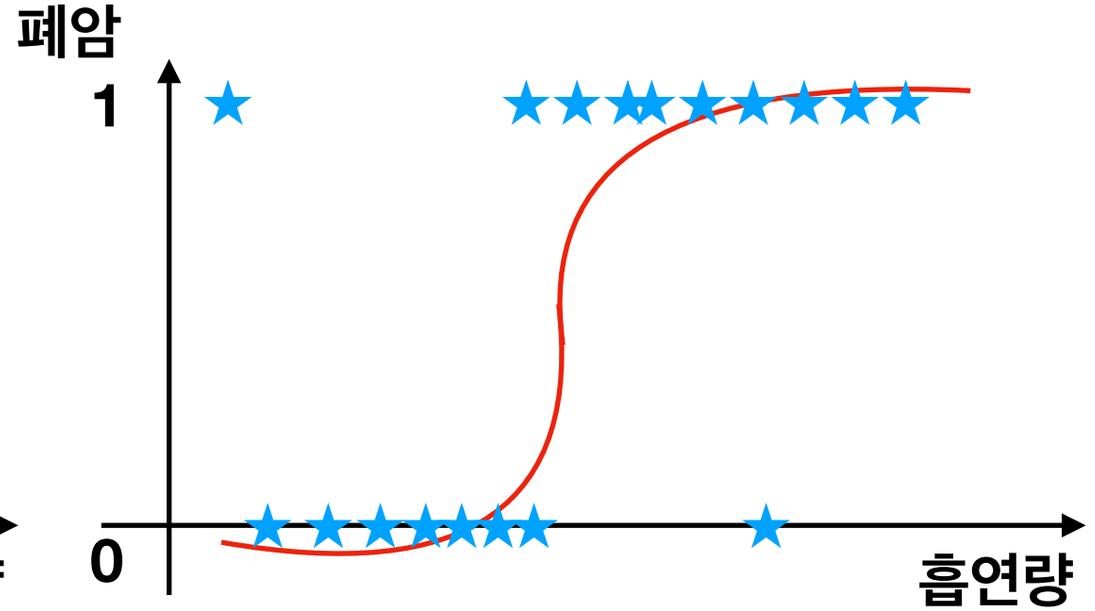
Machine Learning

로지스틱 회귀 (Logistic Regression)

- 범주형 데이터 셋의 경우 선형 회귀로 분류 하기에는 한계가 있음
- Regression이라는 이름과 다르게 Classification으로 보는게 더 적합



$$y = ax + b$$

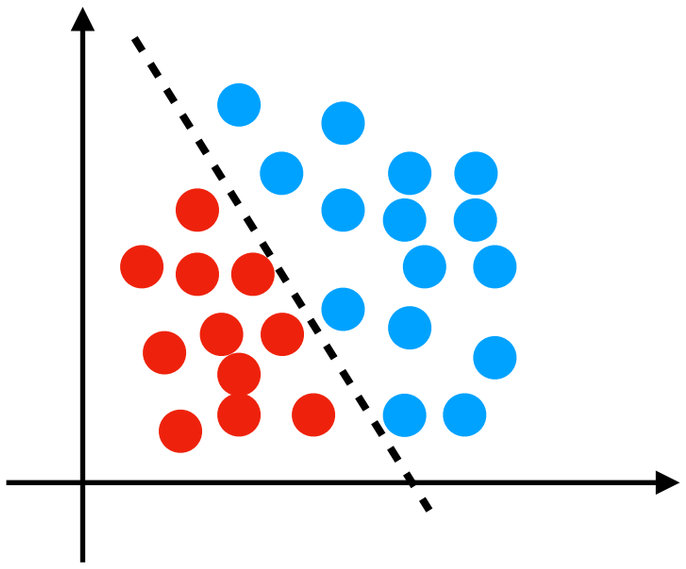


$$y = \frac{1}{1 + e^{-x}}$$

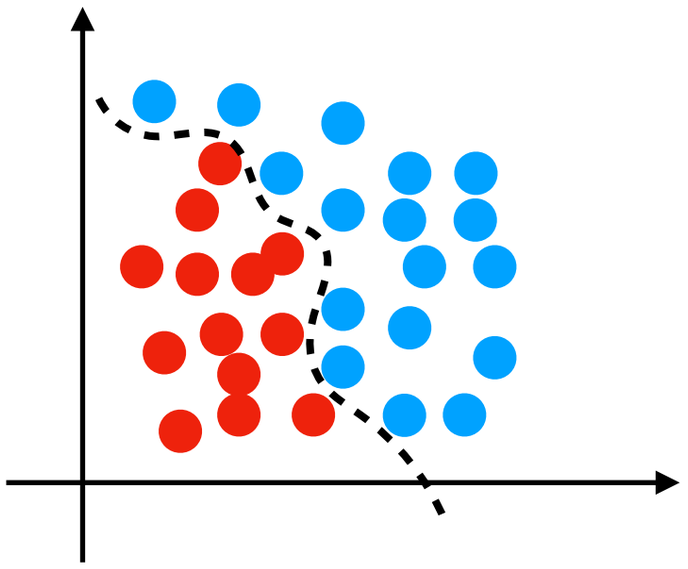
Machine Learning

로지스틱 회귀 (Logistic Regression)의 한계

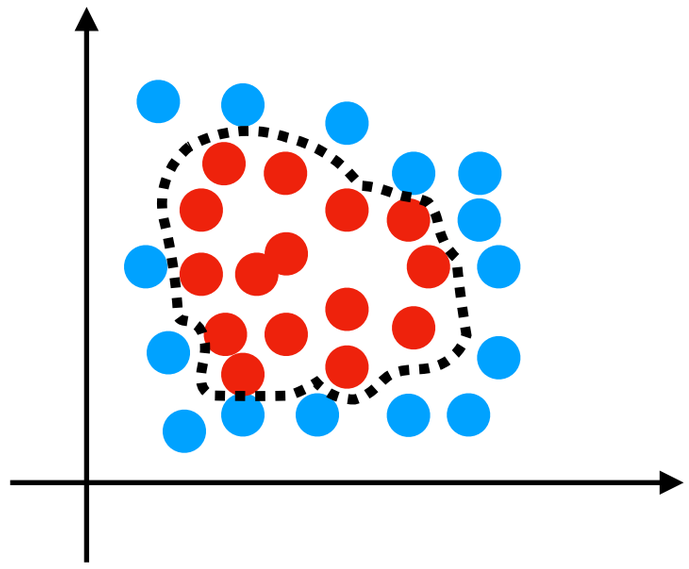
- Non-linearity (비선형성)을 만들기 위해서는 많은 수의 변수 조합이 필요함
  - x1, x2 => x1, x2, x1x2, …
  - x1, x2, x3 => x1, x2, x3, x1x2, x1x3, x2x3, x1x2x3,…
  - 1024x1024픽셀 이미지 데이터의 경우?



$y = a_1x_1 + a_2x_2 + b$



$y = a_1x_1 + a_2x_2 + a_3x_1x_2 + b$

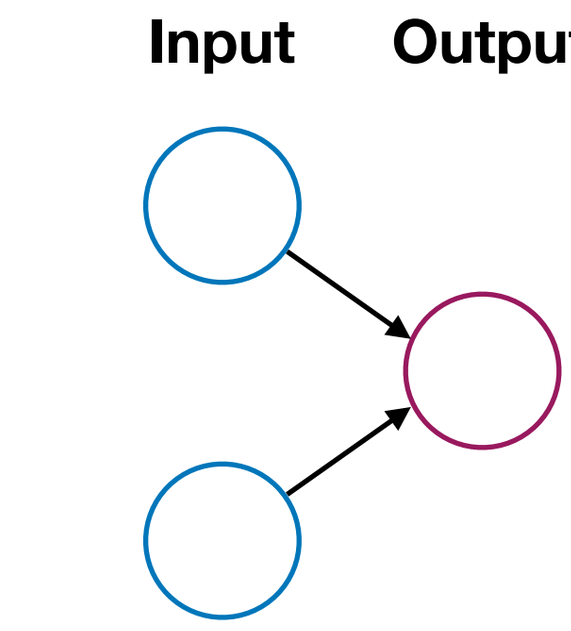


$y = a_1x_1 + a_2x_2 + a_3x_1 \cdots$

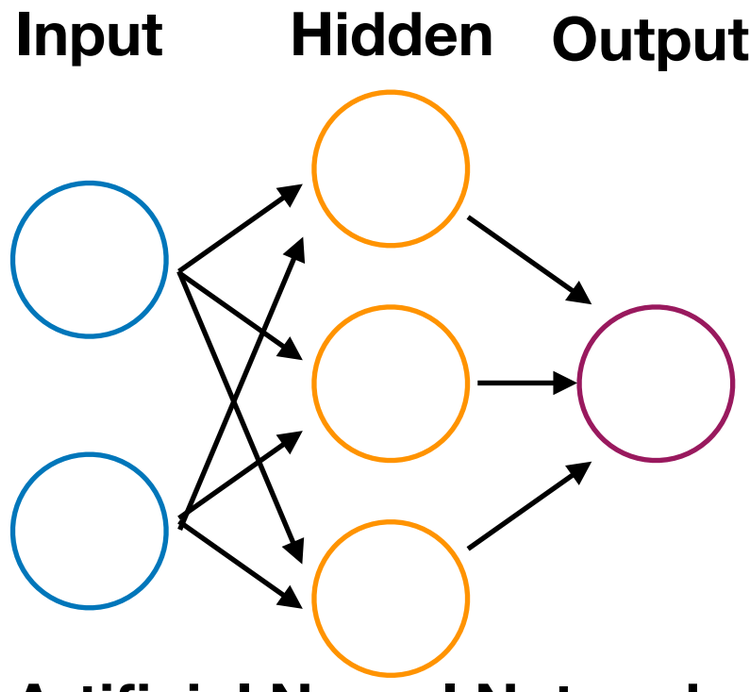
Machine Learning

인공신경망 (Artificial Neural Networks)

- Non-linearity (비선형성)를 제공하기 위해, 변수의 조합이 아닌, 노드의 조합을 이용
  - 각각의 단일 노드 (hidden) 는 하나의 logit과 동일
  - 매 학습 당 독립변수에 곱해지는 파라미터 (weight)를 조정
- 모델이 학습되는 과정에서 값들이 레이어 간의 전파를 통해 이루어 진다고 해서 Feed Forward Neural Networks (FNNs)라고 도 불림



Logistic Regression

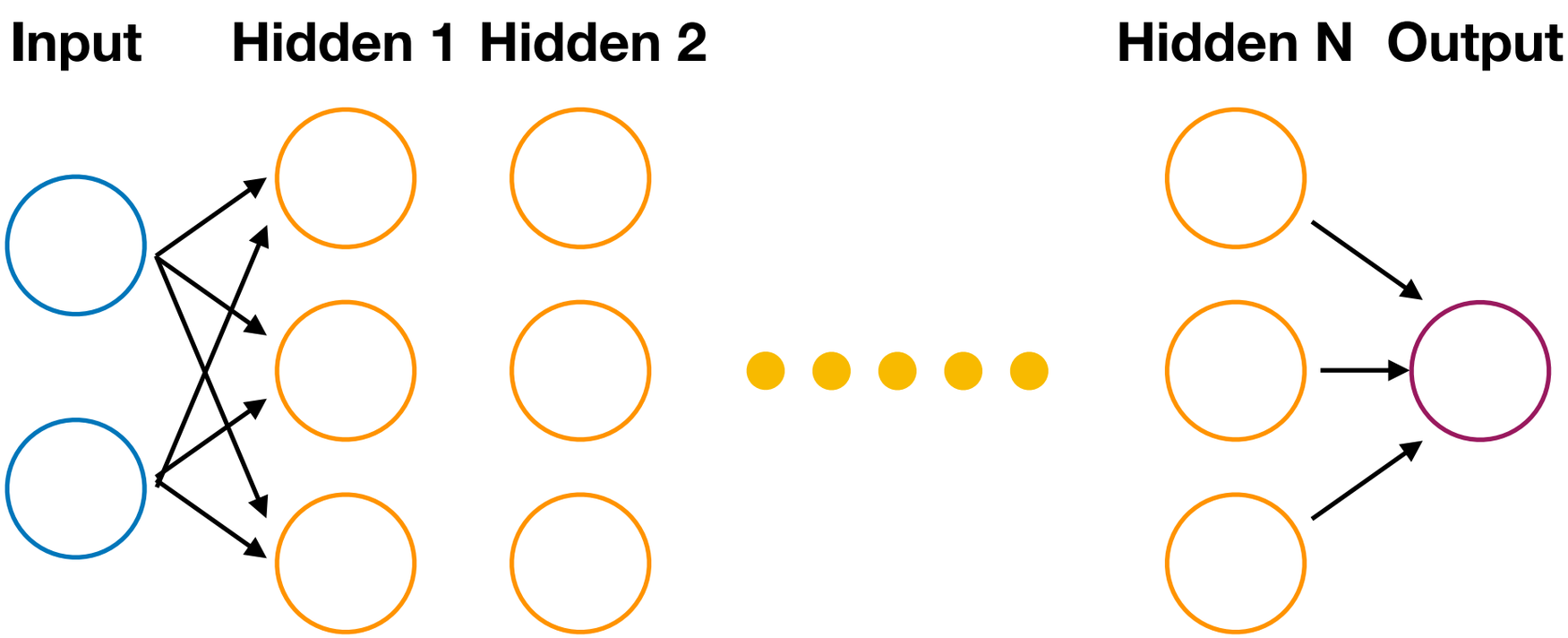


Artificial Neural Networks

Machine Learning

깊은 신경망 (Deep Neural Networks)

- 기존 인공 신경망에 더 많은 Hidden Layer의 수 를 추가해서 깊게 (Deep) 만든 신경망 모델
  - 처음 제시된 시점에 비해 (1970년대) 유명세를 얻기까지 시간이 걸림
  - 효율적인 알고리즘과 컴퓨팅 성능의 향상으로 2010년 대부터 각광을 받기 시작함

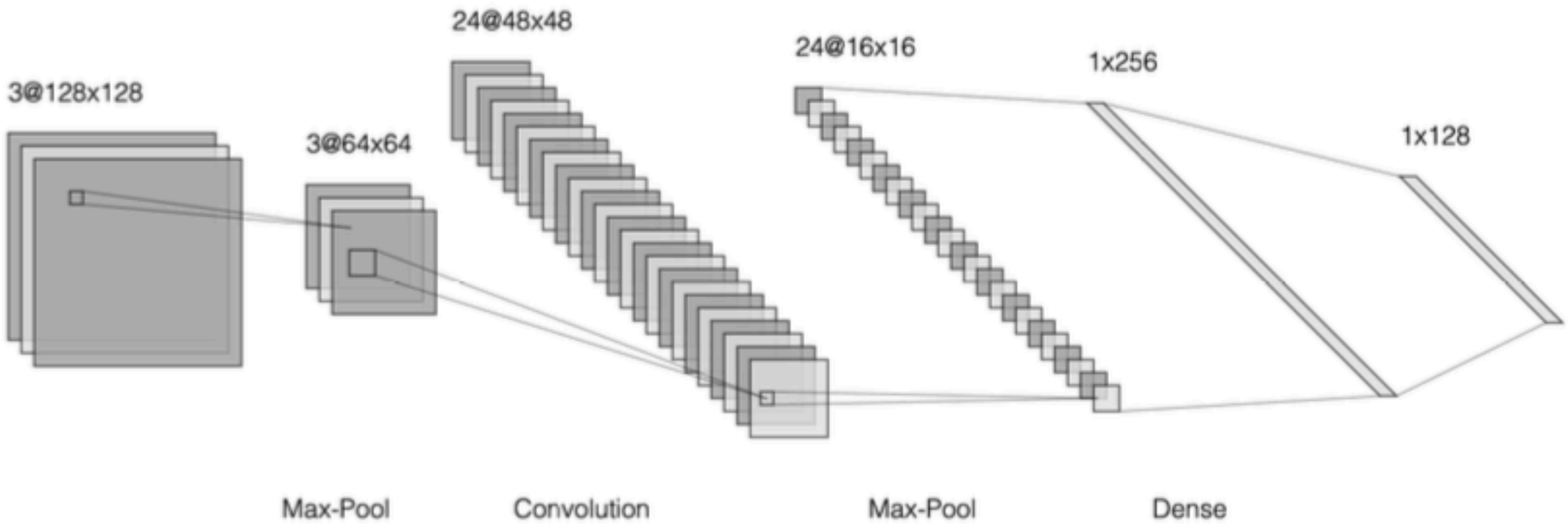


[Deep Neural Networks]

Machine Learning

Convolutional Neural Networks (CNNs)

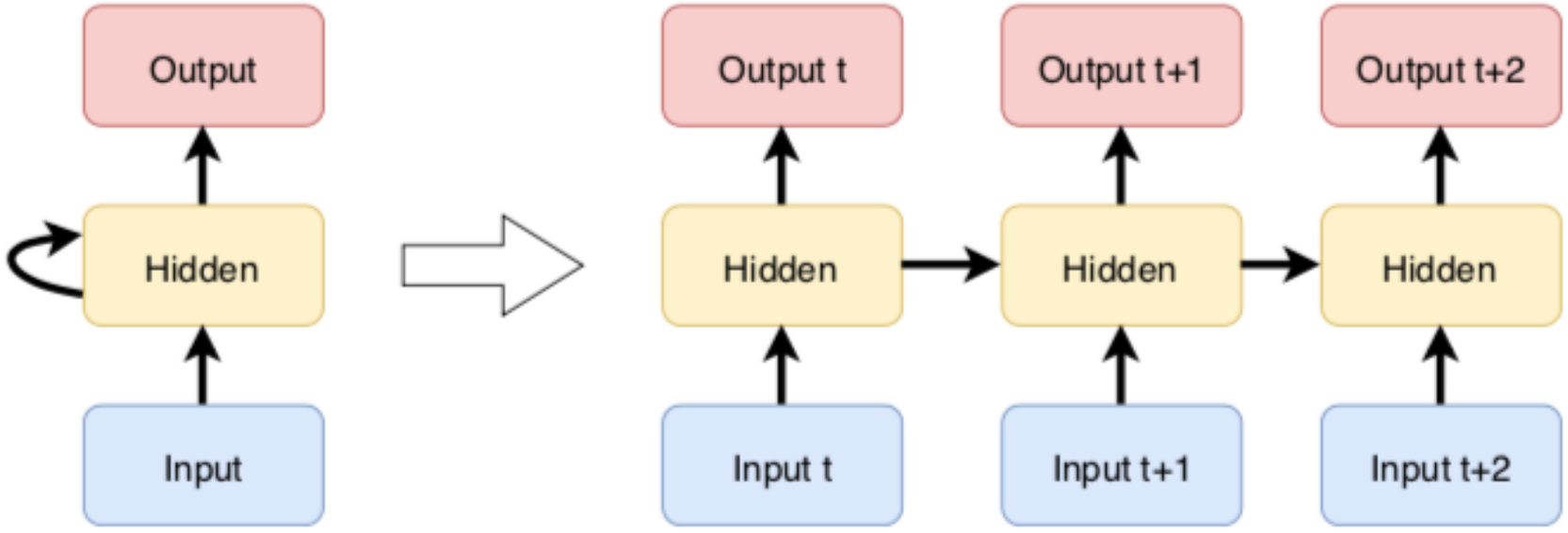
- 이미지 데이터에서 큰 효과를 보여준 모델
- 기존 FNNs의 한계를 보완함
  - 벡터화에 의한 이미지 형태 정보 손실 -> 이미지 원본 형태를 (행렬) 유지한채 학습
  - 벡터화에 의한 기하급수적 모델 파라미터 증가 -> Pooling 레이어로 축소된 이미지 처리



Machine Learning

Recurrent Neural Networks (RNNs)

- 기존 Feed Forward Neural Networks (FNNs) 계열의 경우 시계열 또는 순서를 고려하지 못함
- 순서가 중요한 텍스트 데이터의 경우 RNN류의 모델을 쓰는게 적합
  - LSTM 또는 GRU



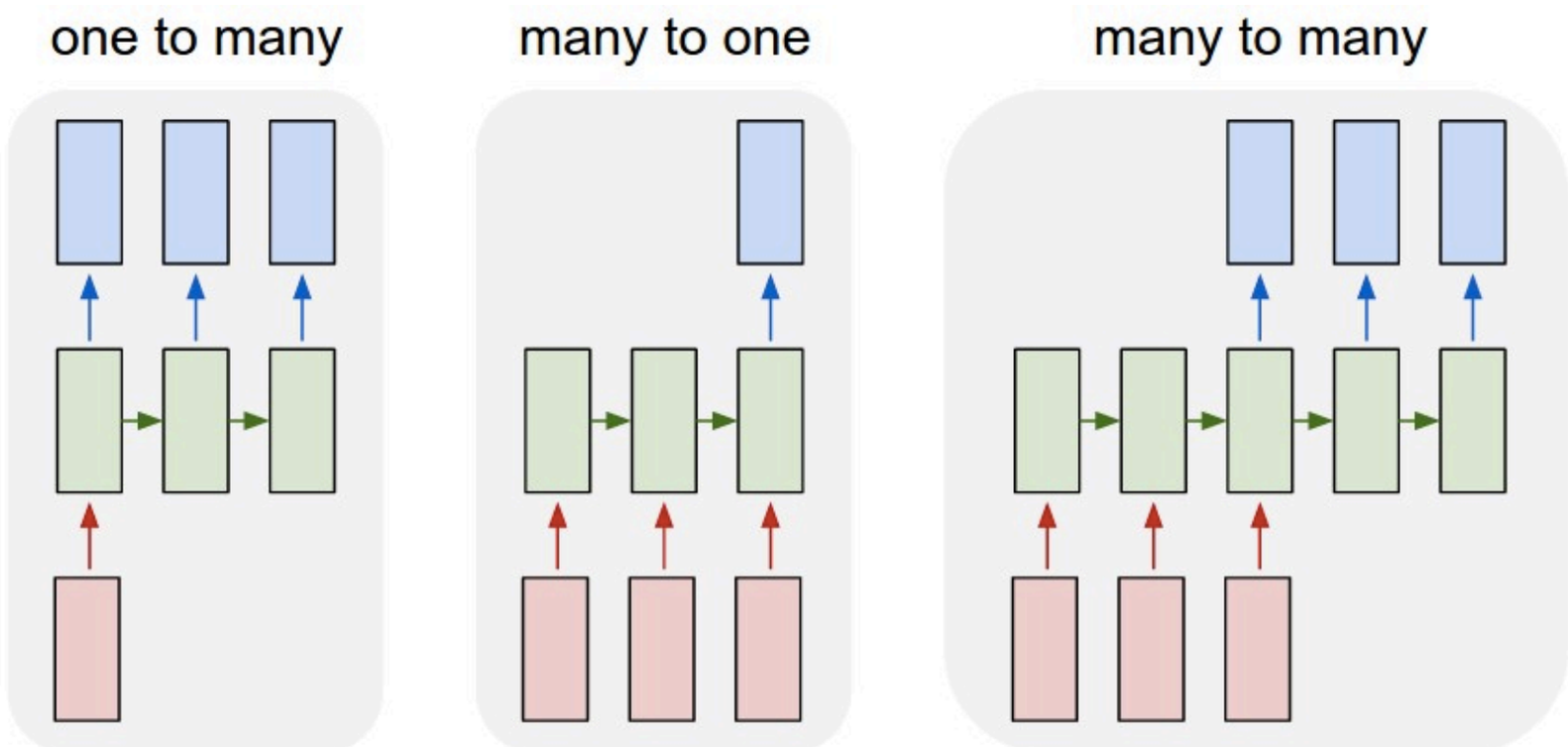
(a) 간소화된 표현 (b) 확장된 표현

[RNNs 의 다이어그램]

Machine Learning

Recurrent Neural Networks (RNNs)

- 목적에 따라 다른 구조를 사용



(a) 이미지 설명 (b) 감성분석 (c) 기계번역, 자동완성

* Source: Andrej Karpathy, The Unreasonable Effectiveness of Recurrent Neural Networks, 2015.05.21, http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Performance Comparison

	FNN	CNN	RNN	LSTM	BILSTM
Training Acc. (%)	58	**87**	80	82	85
Kaggle Acc. (%)	48.7	65.2	58.5	63.7	**65.7**
# of Epochs	34	**6**	161	165	175
Training Time (hours)	**1.5**	1.6	14	15	17.7

Table 3: Result of candidate models

E.O.D