

Introduction to Deep Learning & Neural Networks



AI, ML, DL

ARTIFICIAL INTELLIGENCE

Any technique that enables computers to mimic human behavior



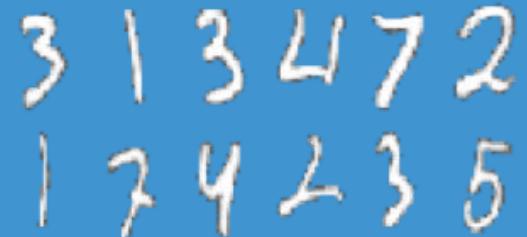
MACHINE LEARNING

Ability to learn without explicitly being programmed



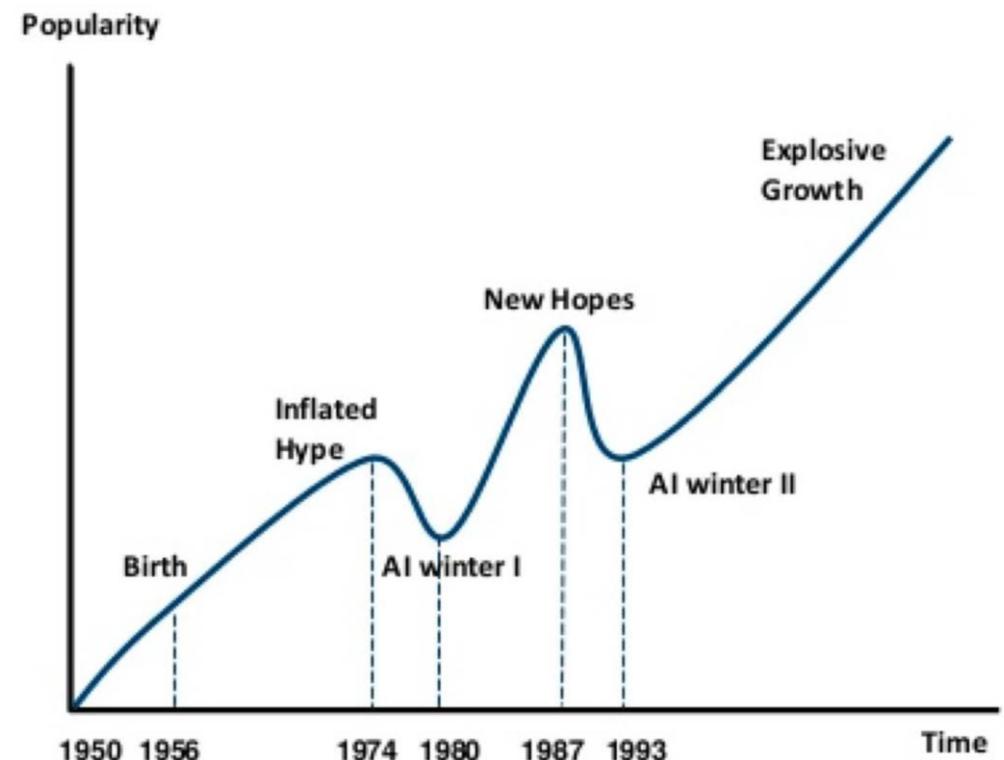
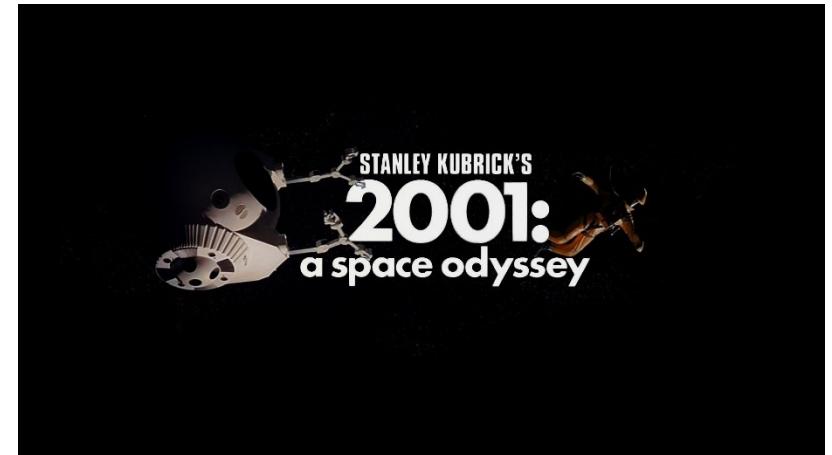
DEEP LEARNING

Extract patterns from data using neural networks



History of Deep Learning(AI)

- 1943: Neural networks
- 1957-62: Perceptron
- 1970-86: Backpropagation, RBM, RNN
- 1979-98: CNN, MNIST, LSTM, Bidirectional RNN
- 2006: “Deep Learning”, DBN
- 2009: ImageNet + AlexNet
- 2014: GANs
- 2016-17: AlphaGo, AlphaZero
- 2017: 2017-19: Transformers



AI vs Human

Deep Blue – 2승 3무 1패



Watson

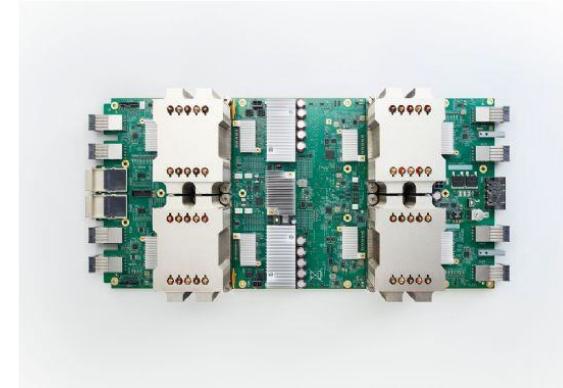


AlphaGo

TPU Server used
against Lee Sedol



AlphaGo Lee – 4승 1패



TPU Board used
against Ke Jie

AlphaGo Master – 3승 0패



Dota 2 & StarCraft 2

Elon Musk's AI beats the world's best Dota 2 players

Staff Writer · 14 August 2017 · 15 Comments

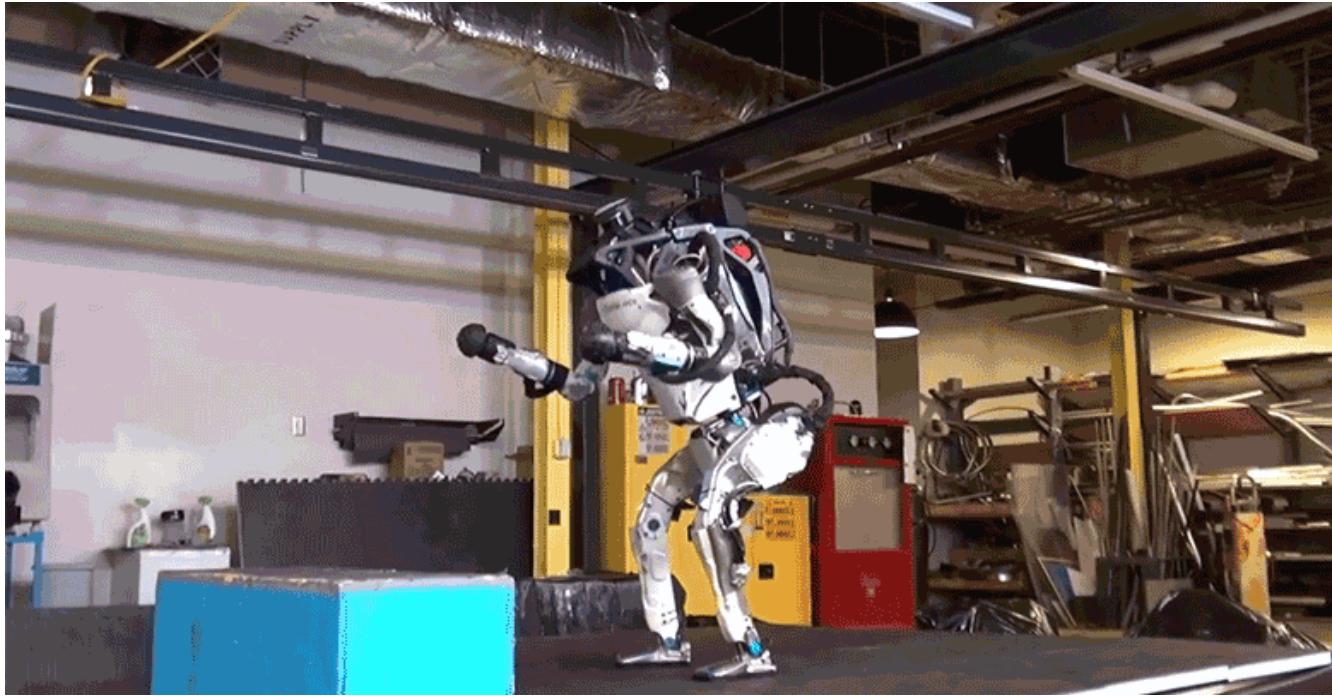
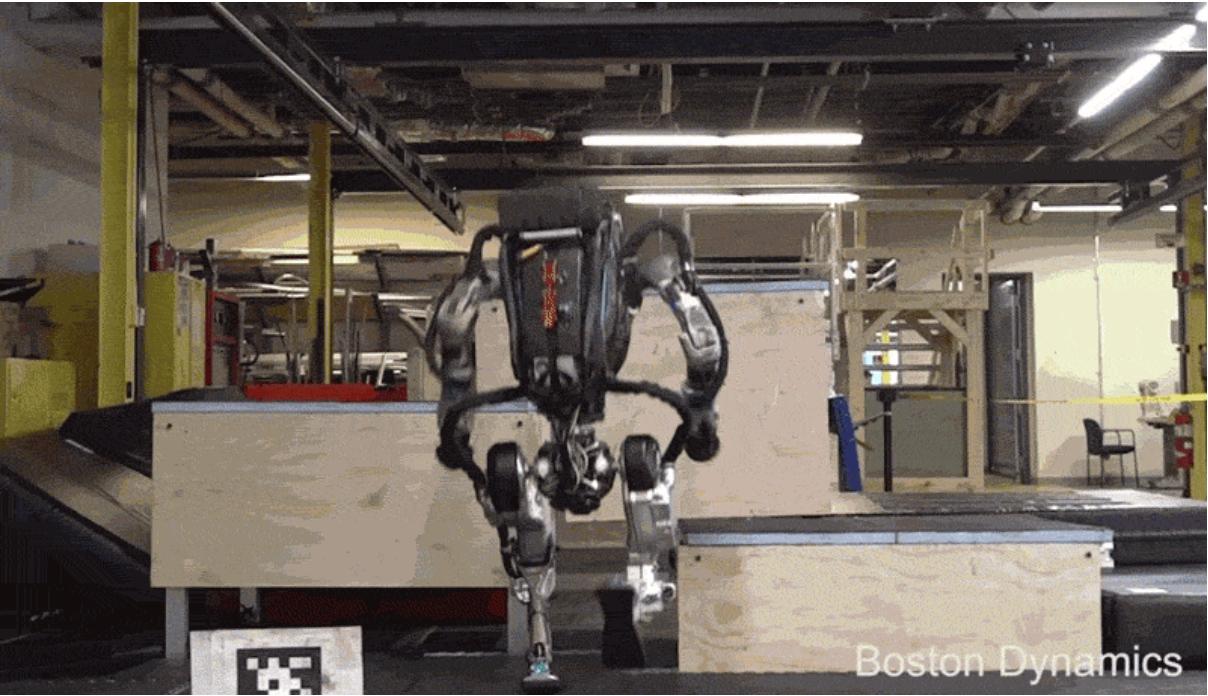


DeepMind's AlphaStar AI wins 10-1 against professional StarCraft II players

Abner Li - Jan. 24th 2019 12:50 pm PT [Twitter](#) [@technacity](#)



Boston Dynamics' Robots



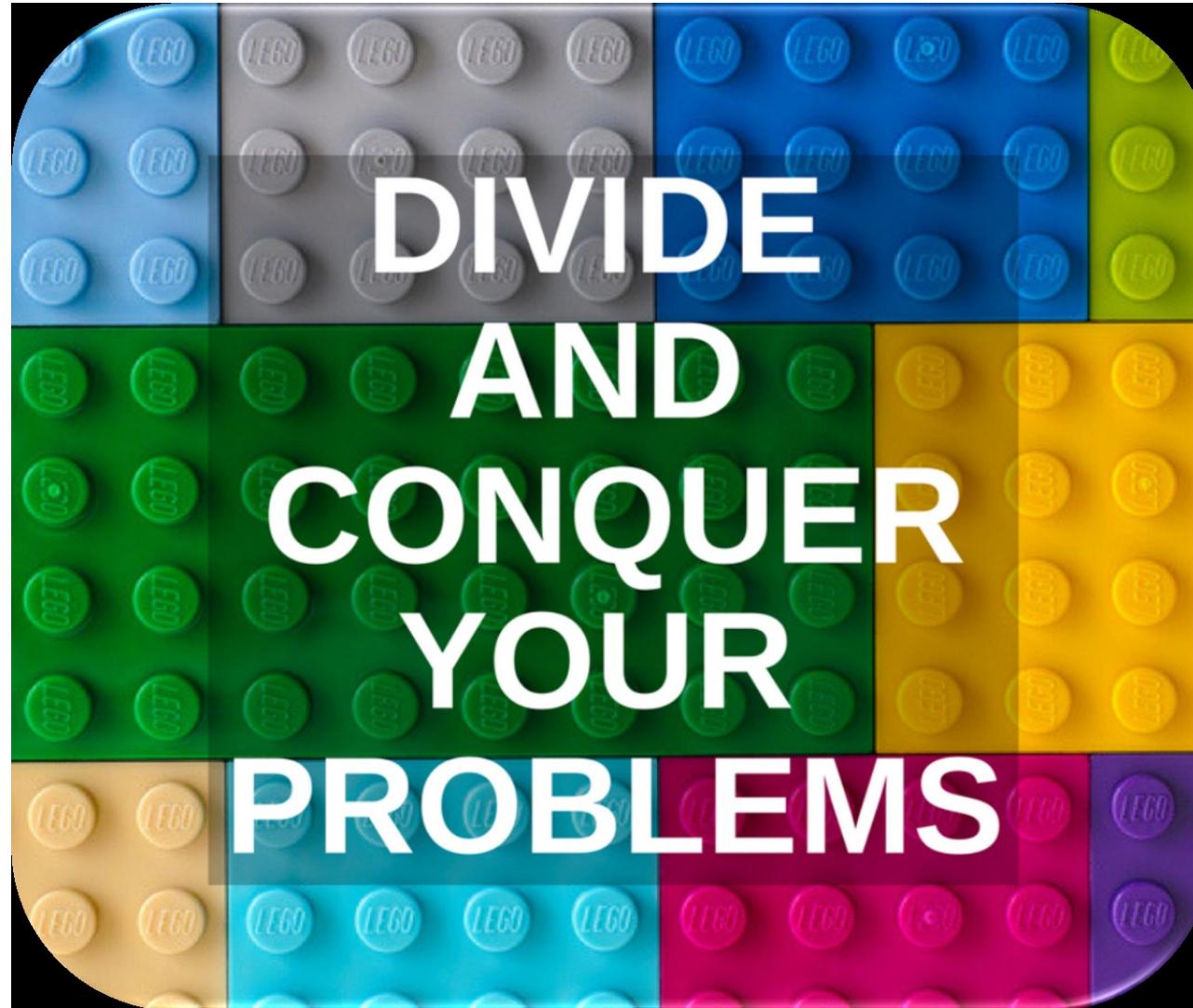
2018 Google Assistant



AI Speaker



Divide and Conquer



What is This?

- The (human) brain is so good at interpreting visual information that the gap between raw data and its semantic interpretation is difficult to assess intuitively:



What is This?

```
array([[[0.03921569, 0.03529412, 0.02352941, 1.      ],
       [0.2509804 , 0.1882353 , 0.20392157, 1.      ],
       [0.4117647 , 0.34117648, 0.37254903, 1.      ],
       ...,
       [0.20392157, 0.23529412, 0.17254902, 1.      ],
       [0.16470589, 0.18039216, 0.12156863, 1.      ],
       [0.18039216, 0.18039216, 0.14117648, 1.      ]],

      [[0.1254902 , 0.11372549, 0.09411765, 1.      ],
       [0.2901961 , 0.2509804 , 0.24705882, 1.      ],
       [0.21176471, 0.2       , 0.20392157, 1.      ],
       ...,
       [0.1764706 , 0.24705882, 0.12156863, 1.      ],
       [0.10980392, 0.15686275, 0.07843138, 1.      ],
       [0.16470589, 0.20784314, 0.11764706, 1.      ]],

      [[0.14117648, 0.12941177, 0.10980392, 1.      ],
       [0.21176471, 0.1882353 , 0.16862746, 1.      ],
       [0.14117648, 0.13725491, 0.12941177, 1.      ],
       ...,
       [0.10980392, 0.15686275, 0.08627451, 1.      ],
       [0.0627451 , 0.08235294, 0.05098039, 1.      ],
       [0.14117648, 0.2       , 0.09803922, 1.      ]],

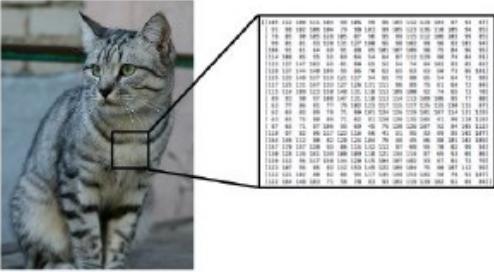
      ...]
```

This is a CAT

<https://github.com/glouuppe/info8010-deep-learning>

What is a CAT?

Viewpoint



Illumination



Deformation



Occlusion



Clutter



Intraclass Variation



Why Learning?

- The automatic extraction of **semantic information** from raw signal is at the core many applications, such as
 - Image recognition
 - Speech processing
 - Natural language processing
 - Robotic control
 - ... and many others
- How can we **write a computer program** that implements that?

Why Learning?

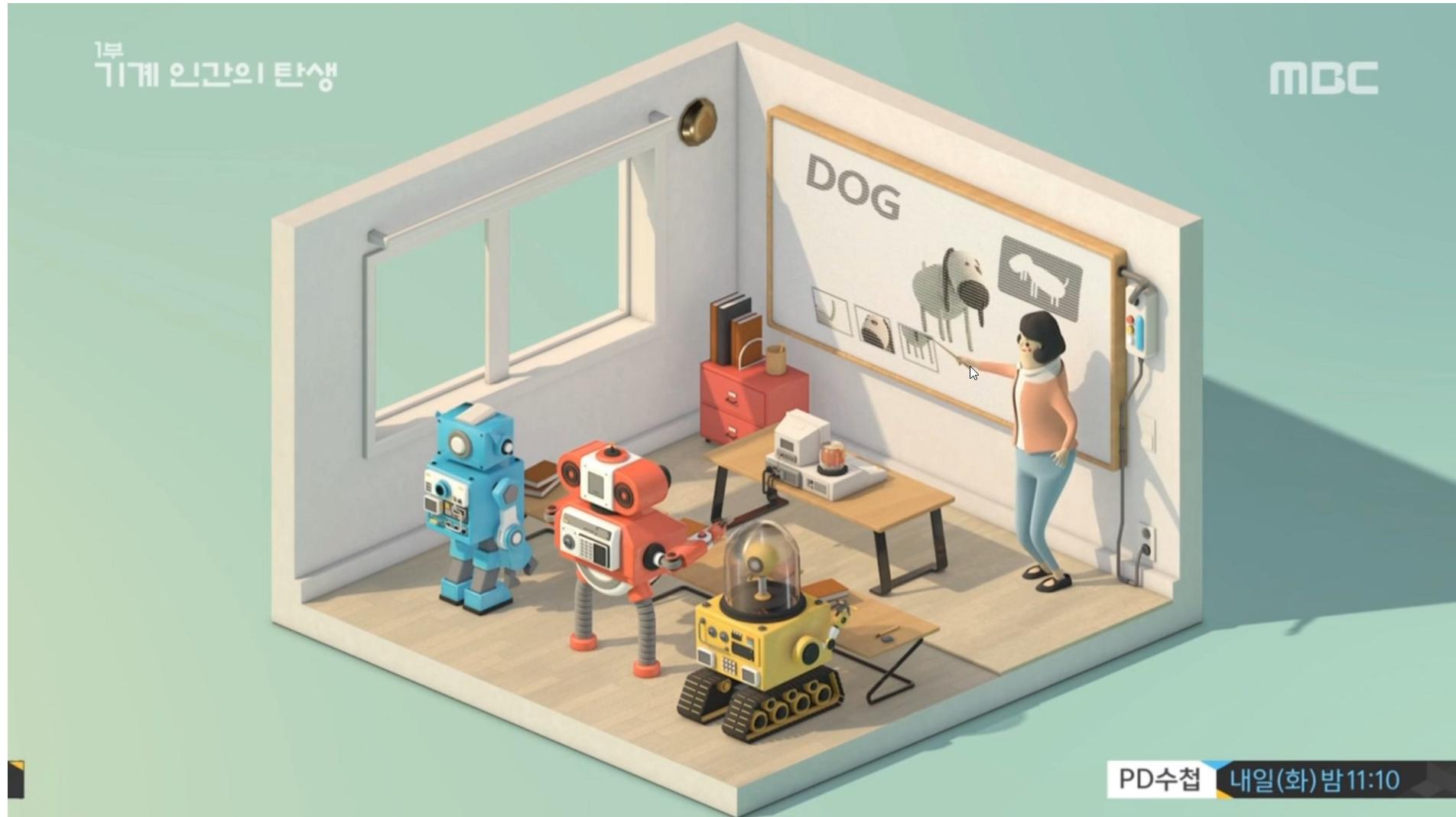


Sheepdog or mop?



Chihuahua or muffin?

어떻게 학습할 것인가?



https://youtu.be/f_uwKZIAeM0

Credit : MBC 기계 인간의 탄생

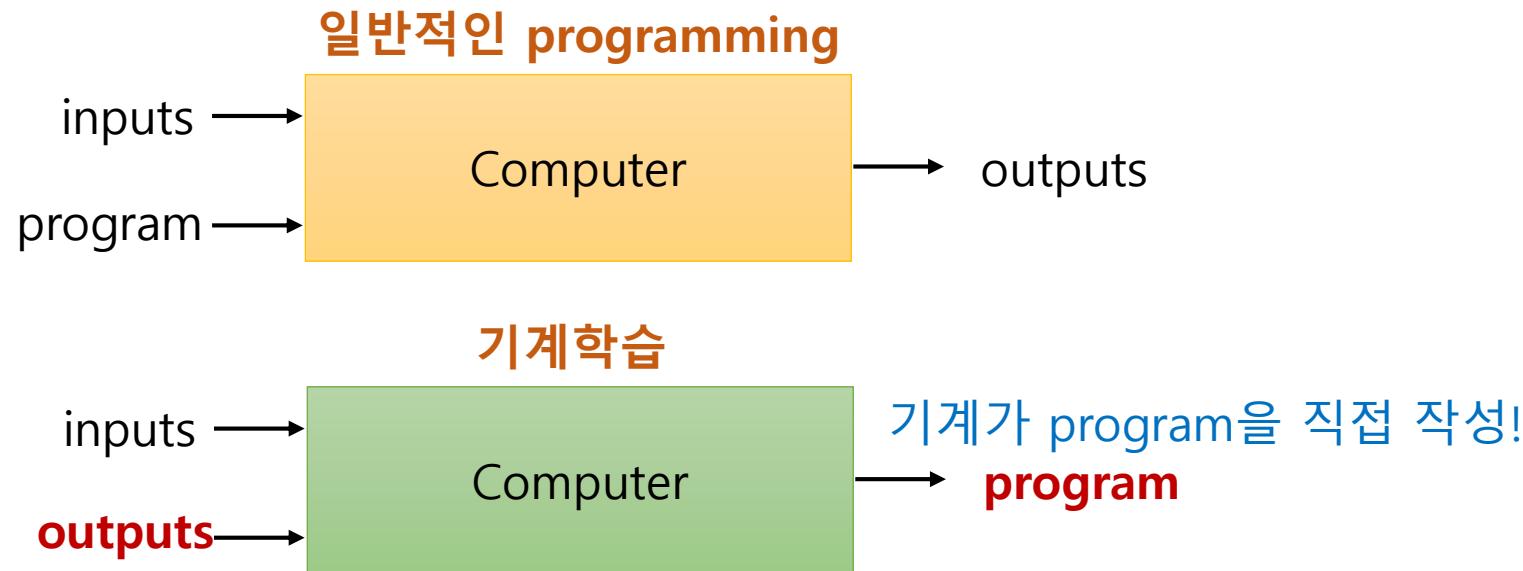
기계학습(Machine Learning)

– Data Driven Approach

- 기계학습(Machine Learning)

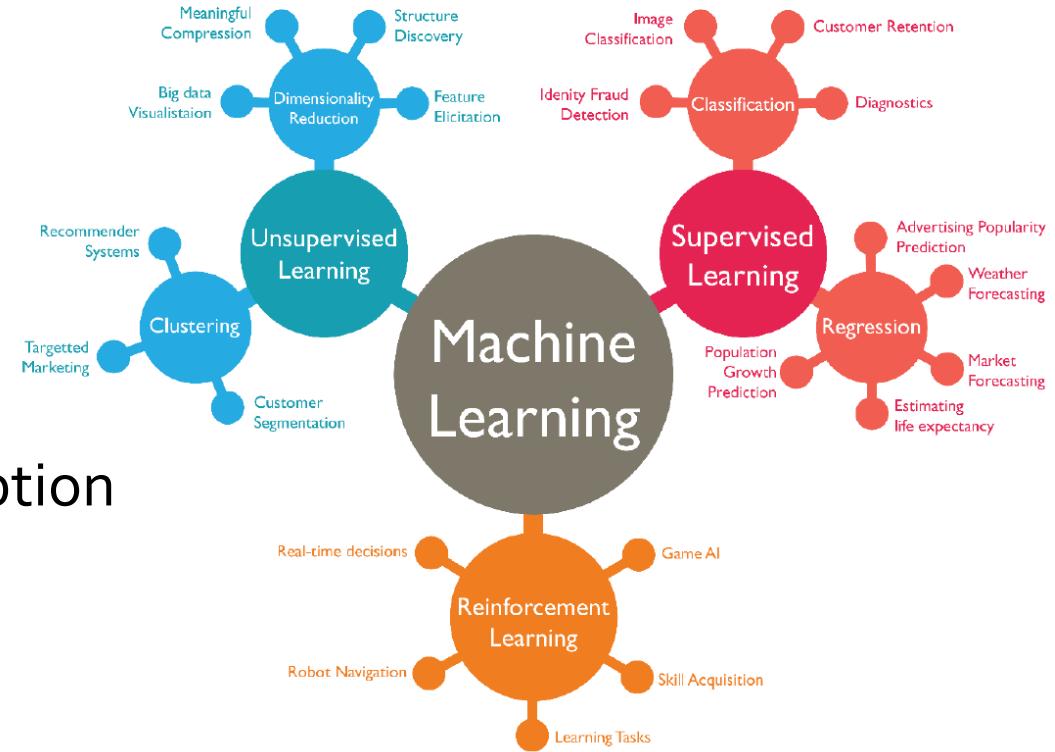
Machine learning is the subfield of [computer science](#) that "gives computers the ability to learn without being explicitly programmed"

기계 학습은 "컴퓨터에 명시적으로 프로그래밍하지 않고 학습 할 수 있는 능력을 부여하는" 컴퓨터 과학의 하위 분야입니다.



Machine Learning의 종류

- Supervised Learning(지도학습)
 - Input 과 labels을 이용한 학습 → function approximator
 - 분류(classification), 회귀(regression)
- Unsupervised Learning(비지도학습)
 - Input만을 이용한 학습 → (shorter) description
 - 군집화(clustering), 압축(compression)
- Reinforcement Learning(강화학습)
 - Trial and error를 통한 학습 → sequential decision making
 - Action selection, policy learning



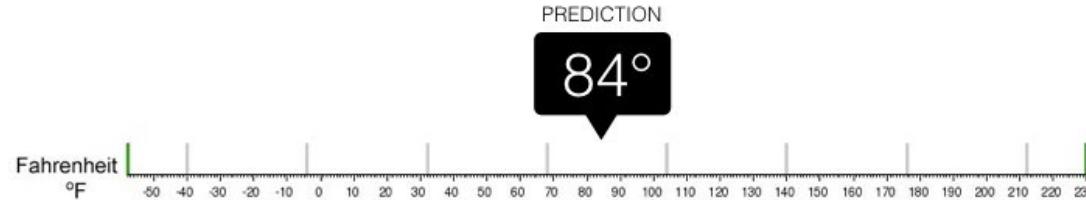
Supervised Learning

– Regression vs Classification



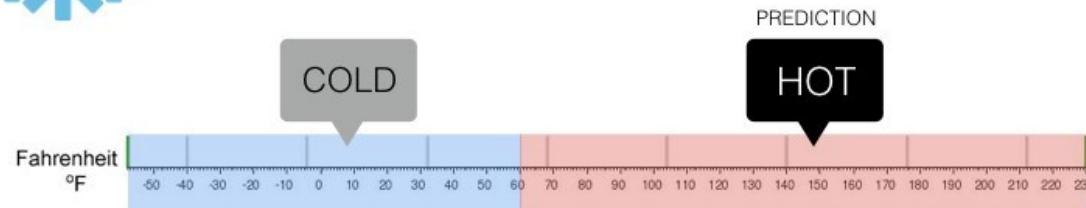
Regression

What is the temperature going to be tomorrow?



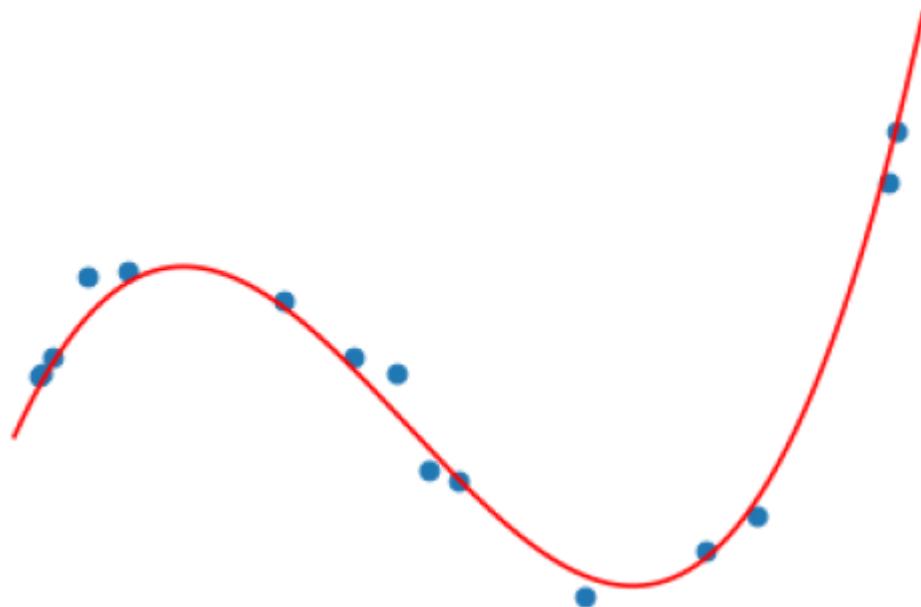
Classification

Will it be Cold or Hot tomorrow?

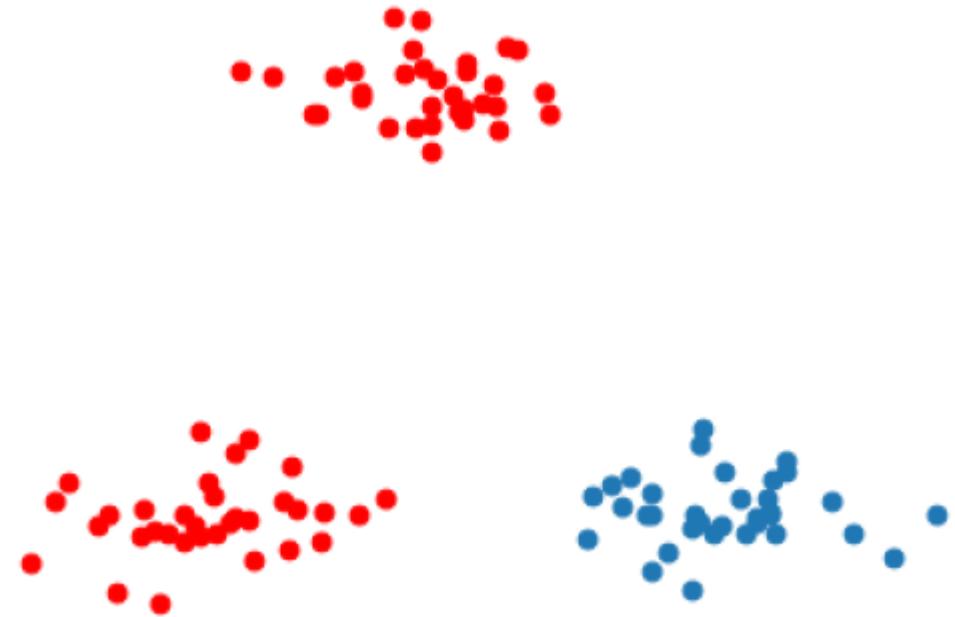


Supervised Learning

– Regression vs Classification



Regression aims at estimating relationships among
(using continuous) variables



Classification consists in identifying a decision boundary
between objects of distinct classes

Supervised Learning

– Regression vs Classification

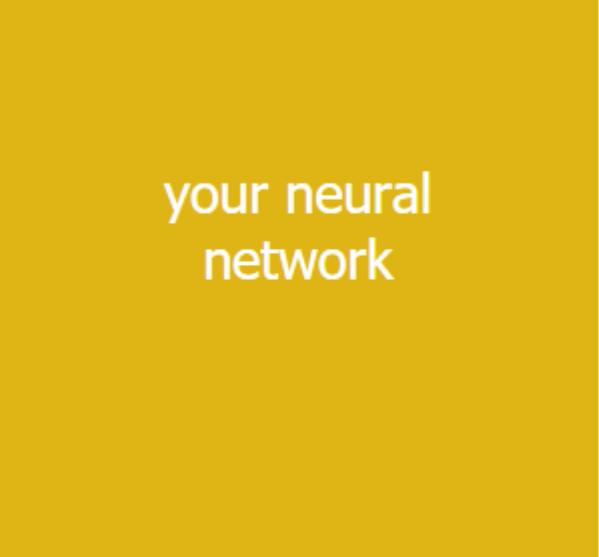
- **Classification:** Given $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^p \times \{1, \dots, C\}$, for $i = 1, \dots, N$, we want to estimate for any new \mathbf{x} ,

$$\arg \max_y P(Y = y | X = \mathbf{x}).$$

- **Regression:** Given $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^p \times \mathbb{R}$, for $i = 1, \dots, N$, we want to estimate for any new \mathbf{x} ,

$$\mathbb{E}[Y | X = \mathbf{x}] .$$

Supervised Learning



your neural
network



YIQ
formula

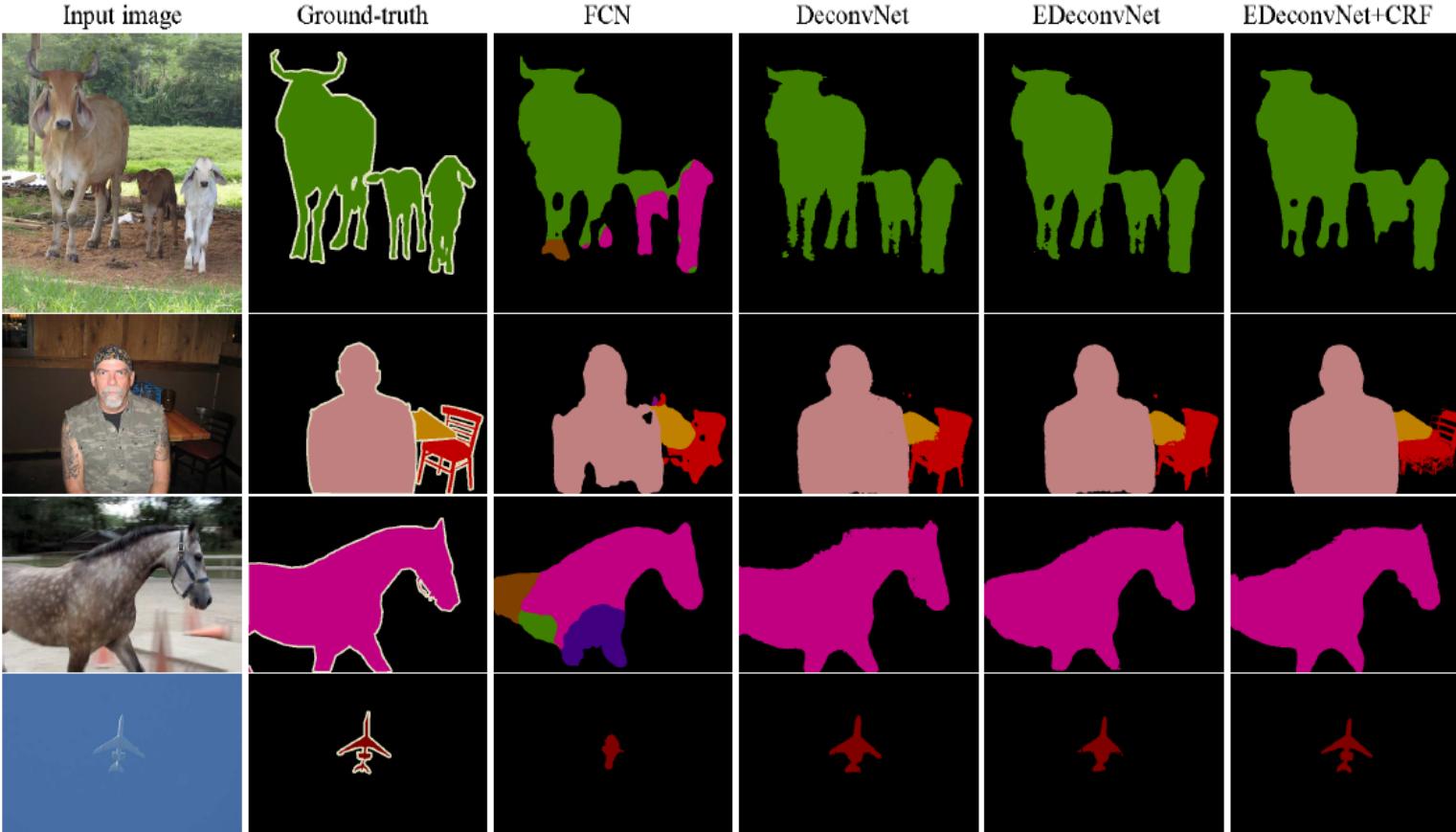
code ►

code ►

Object Detection

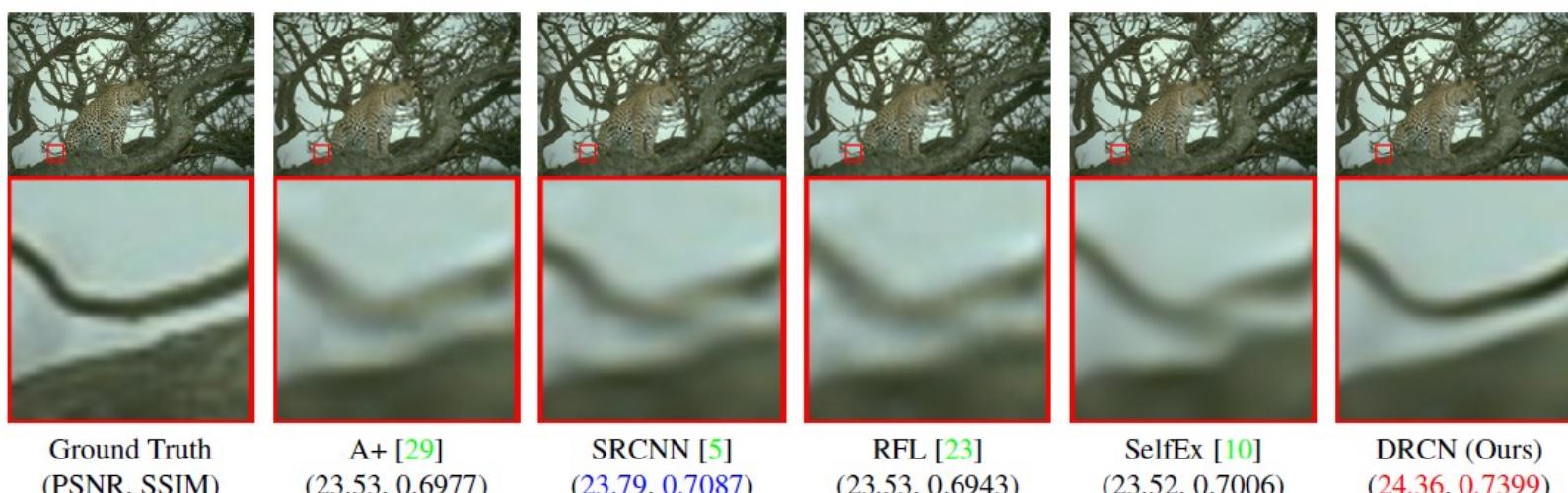
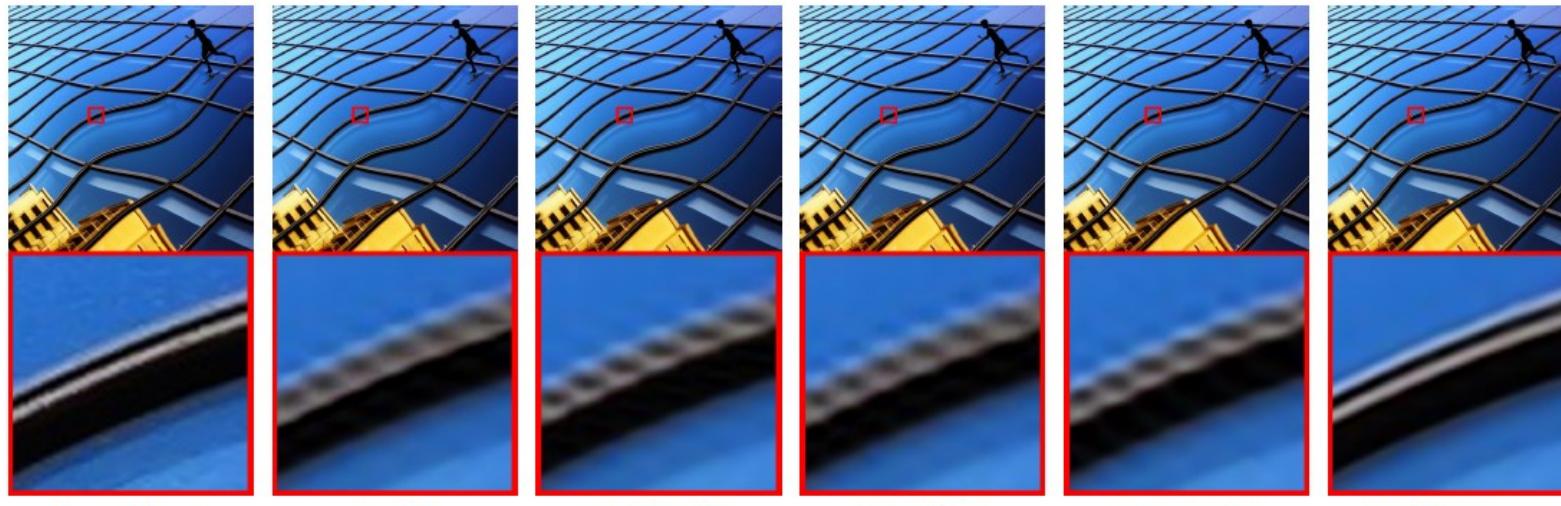


Segmentation



(a) Examples that our method produces better results than FCN [19].

Super Resolution



"Deeply-Recursive Convolutional Network for Image Super-Resolution"

Machine Translation

영어 한국어 독일어 언어 감지 ▾

한국어 프랑스어 영어 ▾ 번역하기

옛날에 백조 한 마리가 살았다.

x

17/5000

◀ 🔍 ⌂ ▾

Once upon a time a swan lived.

☆ 🔍 ◀ ↵

수정 제안하기

yesnal-e baegjo han maliga sal-assda.

영어 한국어 독일어 언어 감지 ▾

한국어 프랑스어 영어 ▾ 번역하기

옛날에 백조 한 마리가 살았습니다.

x

19/5000

◀ 🔍 ⌂ ▾

The 100,000,000,000,001 lived long ago. ✓

☆ 🔍 ◀ ↵

수정 제안하기

yesnal-e baegjo han maliga sal-assseubnida.

Image Captioning



a group of people standing around a room with remotes
logprob: -9.17



a young boy is holding a baseball bat
logprob: -7.61



a cow is standing in the middle of a street
logprob: -8.84



a baby laying on a bed with a stuffed bear
logprob: -8.66



a young boy is holding a baseball bat
logprob: -7.65



a woman holding a teddy bear in front of a mirror
logprob: -9.65

Visual QnA

Q: What is the boy holding?



DPPnet: **surfboard**

Q: What is the animal doing?



DPPnet: **resting** (relaxing)



DPPnet: **bat**



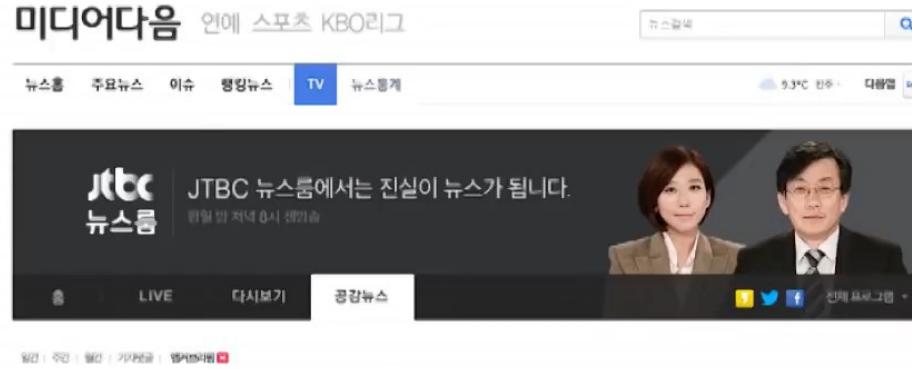
DPPnet: **swimming** (fishing)

Auto Speech Recognition



Text to Speech

kakao



Unsupervised Learning

- 아래 음식을 둘로 분류하면?



(1)



(2)



(3)



(4)



(5)



(6)



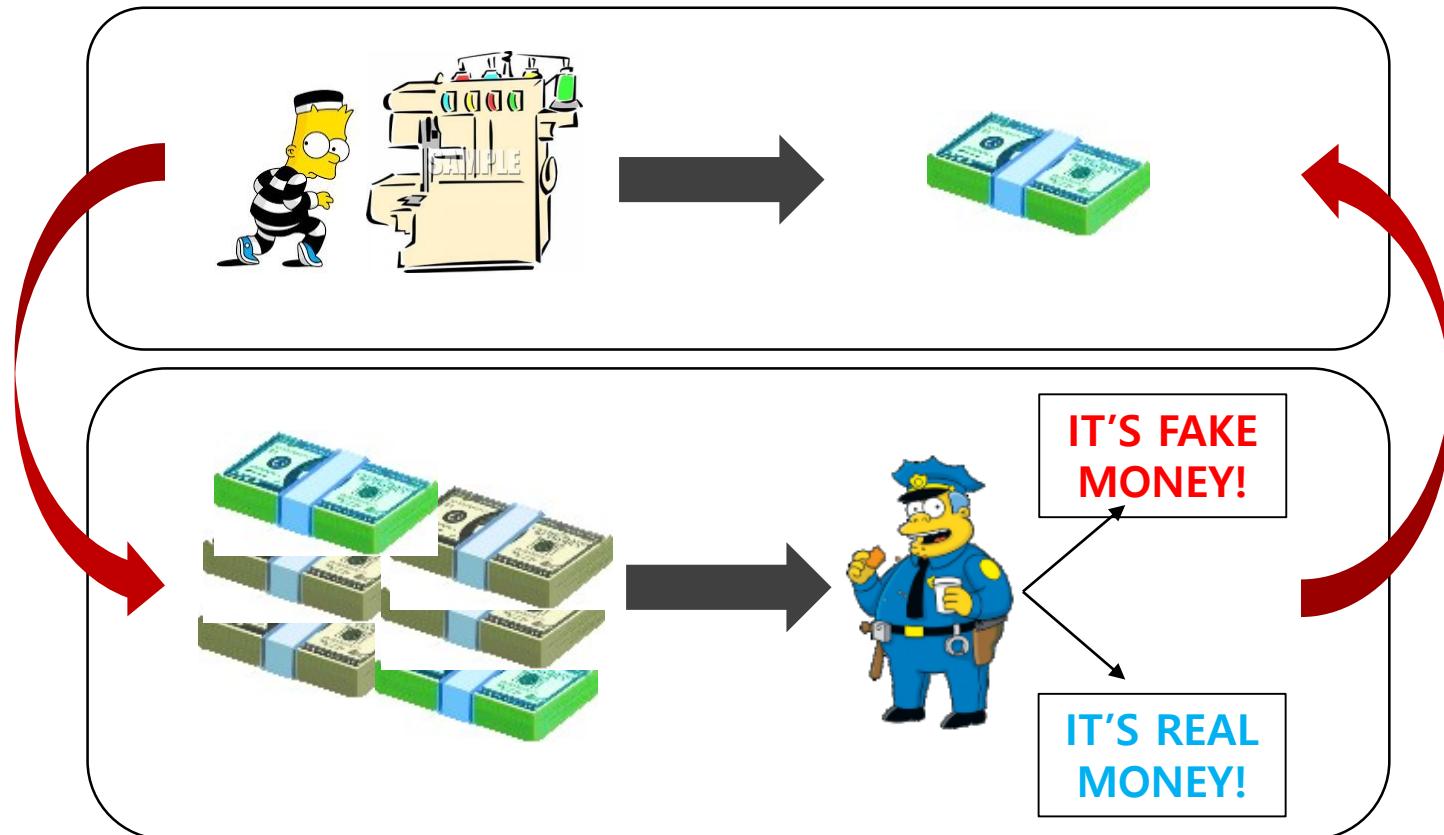
(7)



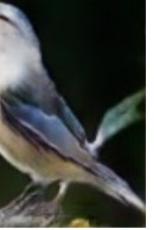
(8)

Generative Adversarial Network

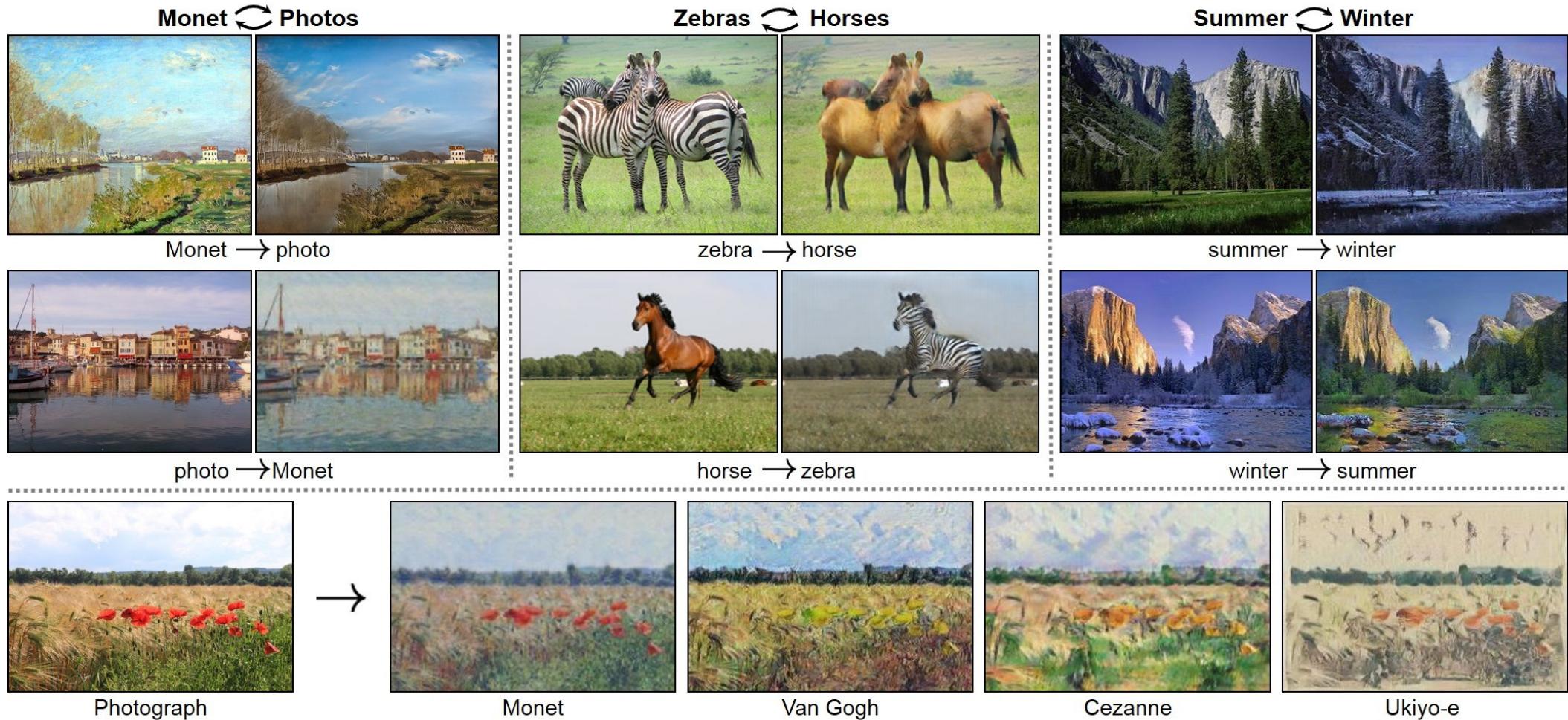
- Counterfeitors vs Police Game



Text to Image Generation

| | | | | | | | |
|------------------------------|---|--|---|---|---|---|---|
| Text description | This bird is red and brown in color, with a stubby beak | The bird is short and stubby with yellow on its body | A bird with a medium orange bill white body gray wings and webbed feet | This small black bird has a short, slightly curved bill and long legs | with varying shades of brown with white under the eyes | bird with a black crown and a short black pointed beak | has a white breast, light grey head, and black wings and tail |
| 64x64 GAN-INT-CLS [22] |  |  |  |  |  |  |  |
| 128x128 GAWWN [20] |  |  |  |  |  |  |  |
| 256x256 StackGAN |  |  |  |  |  |  |  |

CycleGAN

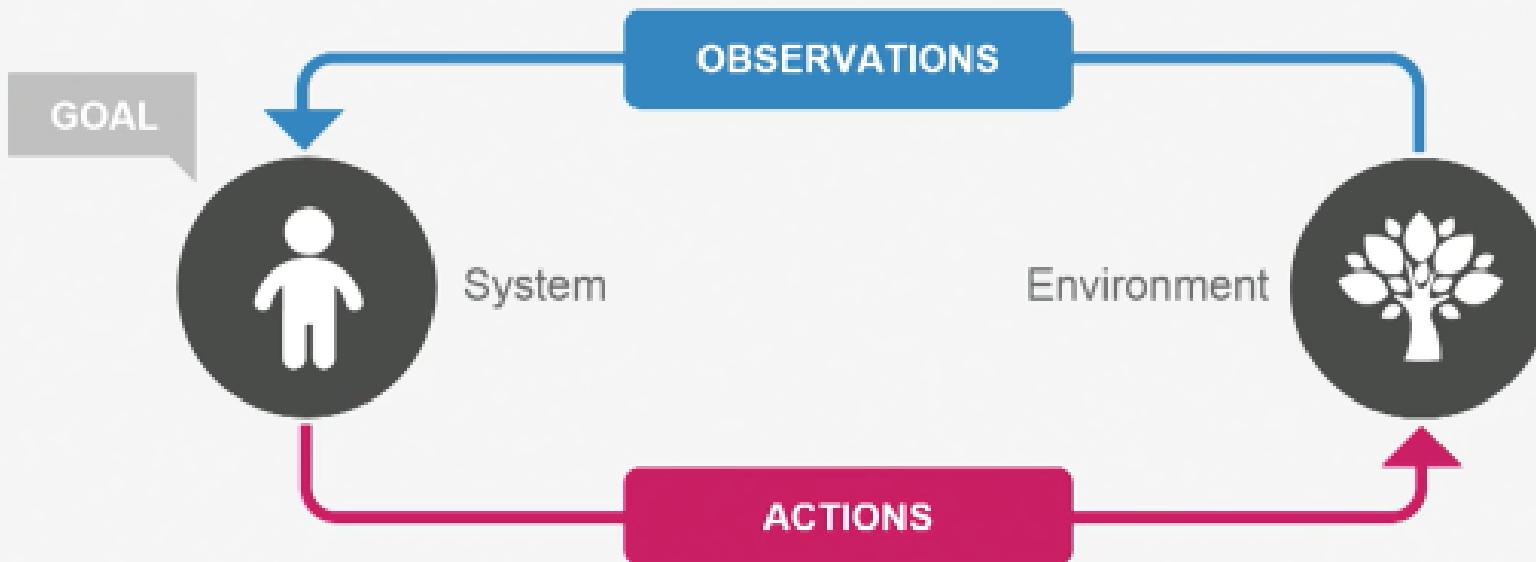


Video Generation

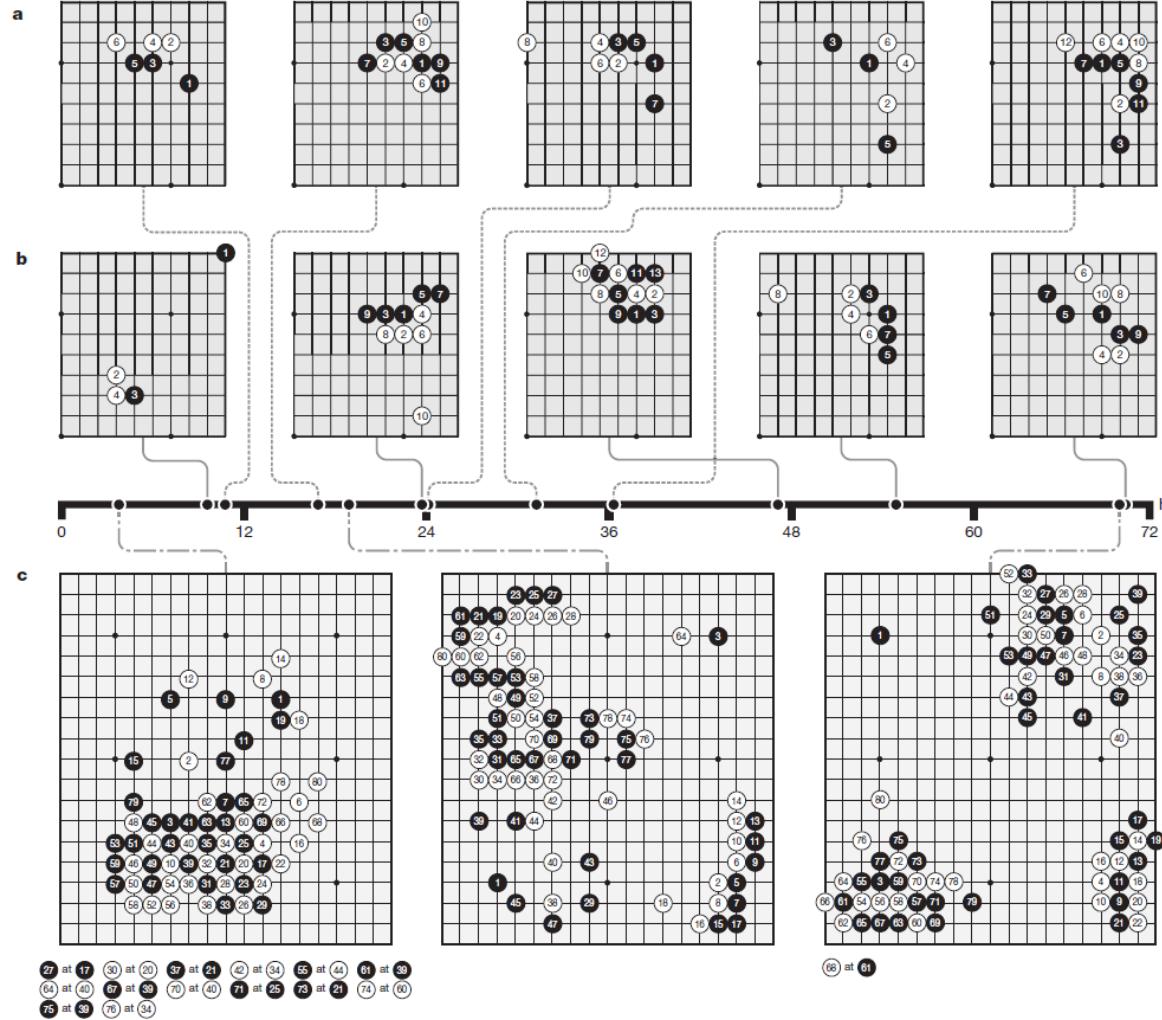


Reinforcement Learning

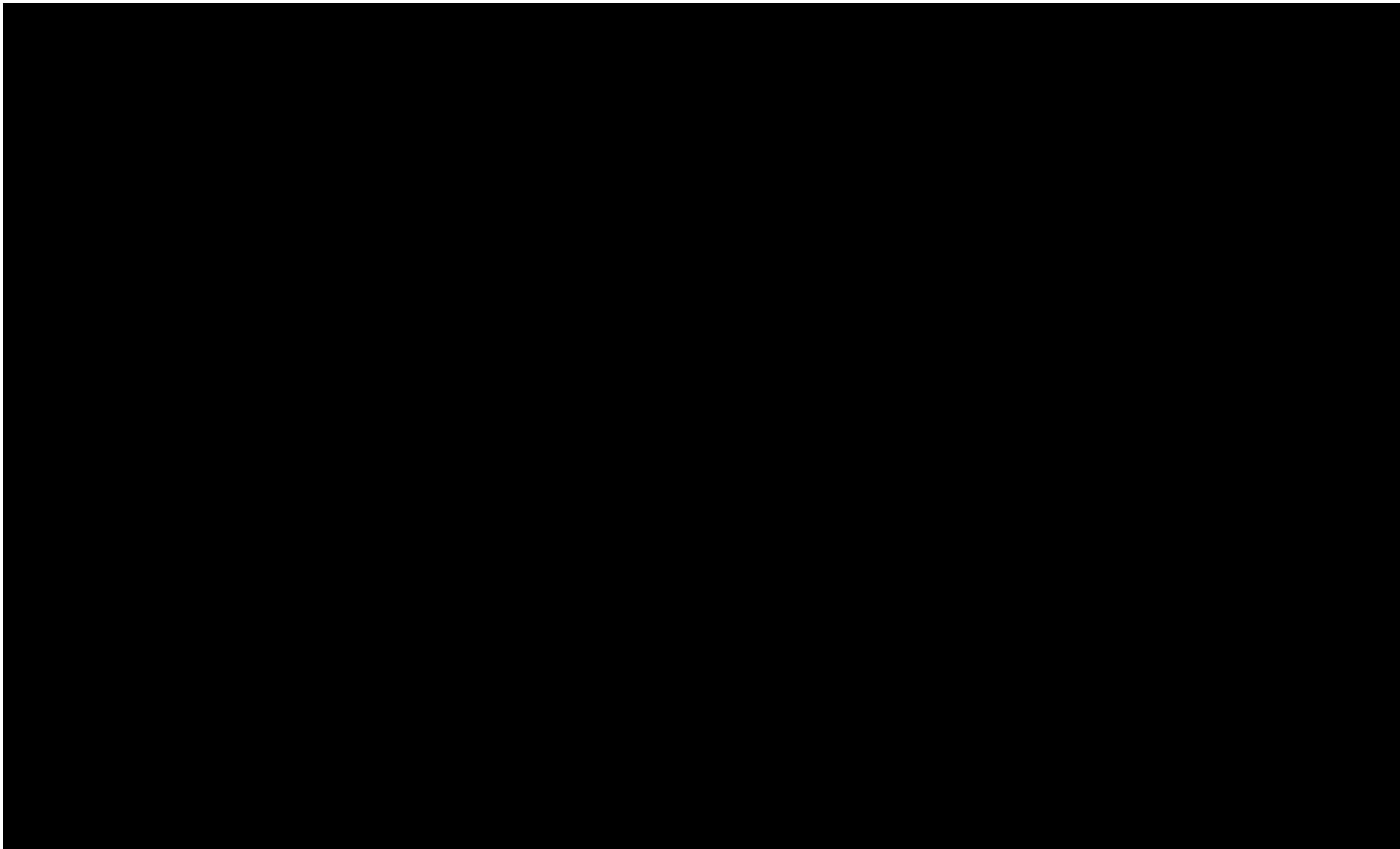
Reinforcement Learning Framework



AlphaGo Zero



Atari Games



Atari Games

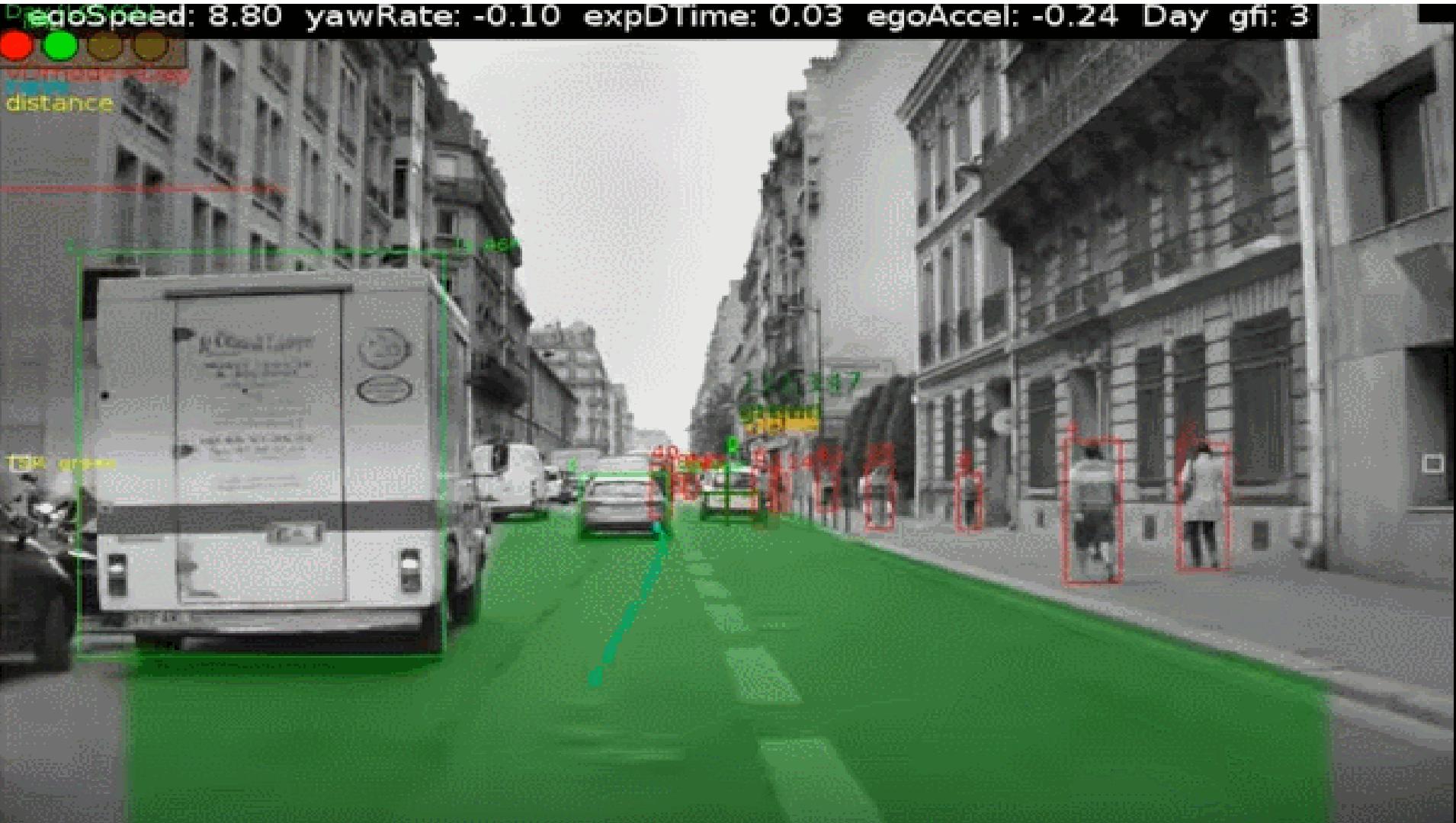
Space Invaders



Other Games

- Super Mario World
 - https://youtu.be/L4KBBAwF_bE
 - <https://youtu.be/qv6UVOQoF44>
- Cookie Run
 - <https://youtu.be/exXD6wJLJ6s>
- Starcraft I
 - <https://youtu.be/GgkmJDjeJtw>
- GTA
 - <https://youtu.be/X4u2DCOLolg>

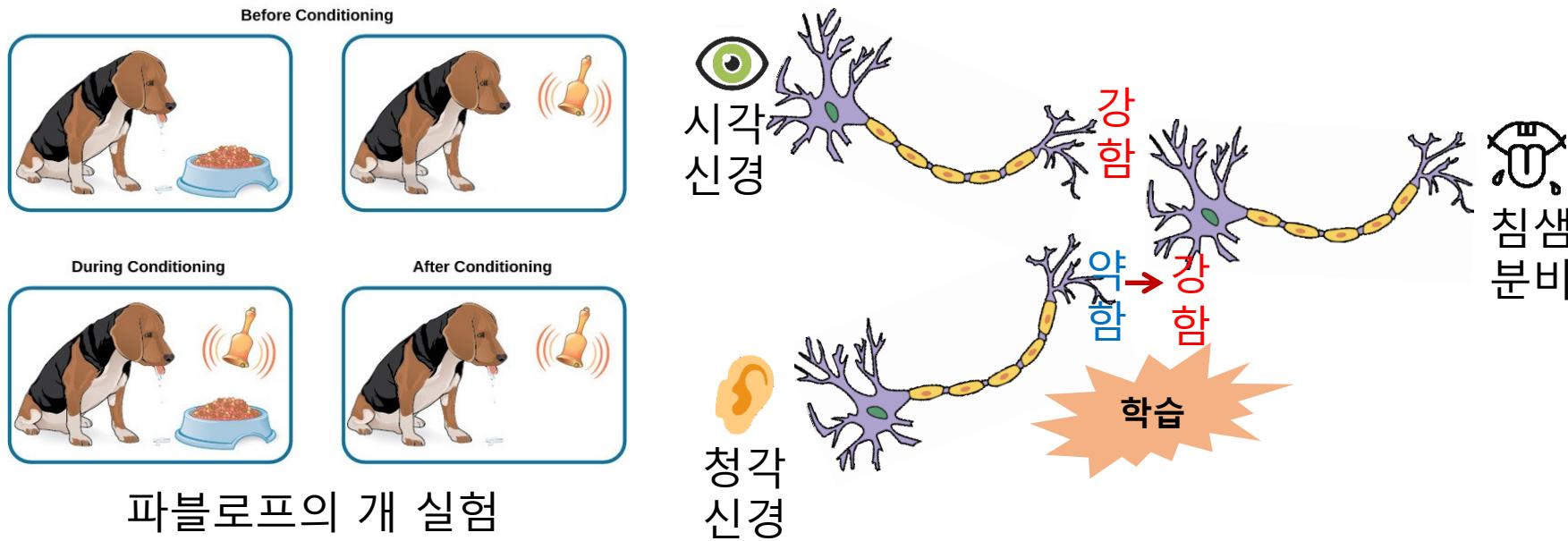
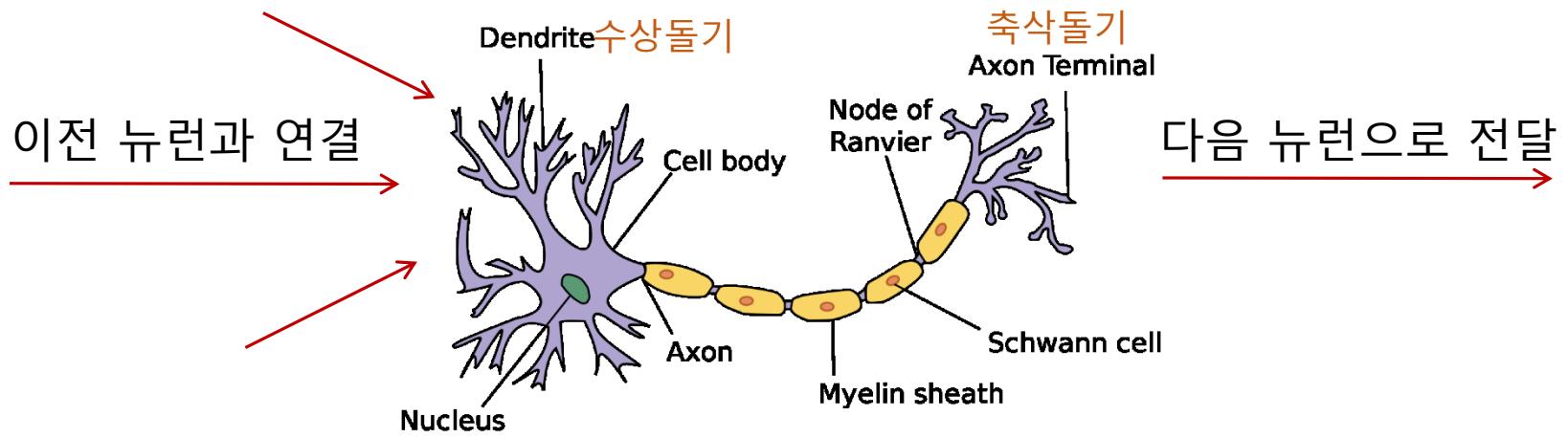
Autonomous Driving



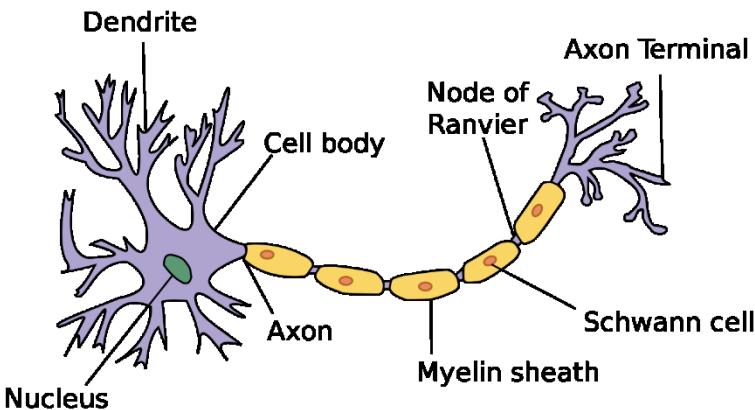
Quiz

- □와 △에 들어갈 정수는?
 - $3 \times \square + 2 \times \triangle = 1$
 - $1 \times \square + 4 \times \triangle = -3$
 - $5 \times \square + 5 \times \triangle = 0$
 - $8 \times \square + 3 \times \triangle = 5$
- $\square = 1, \triangle = -1$
- $(3, 2), (1, 4), (5, 5), (8, 3)$ 은 input data, $1, -3, 0, 5$ 는 label이다
- □와 △를 weight라고 하며 이 weight 값을 기계가 스스로 학습을 통해 찾아내도록 하는 것이 neural network를 이용한 기계학습이 하는 일

뉴런과 사람의 학습



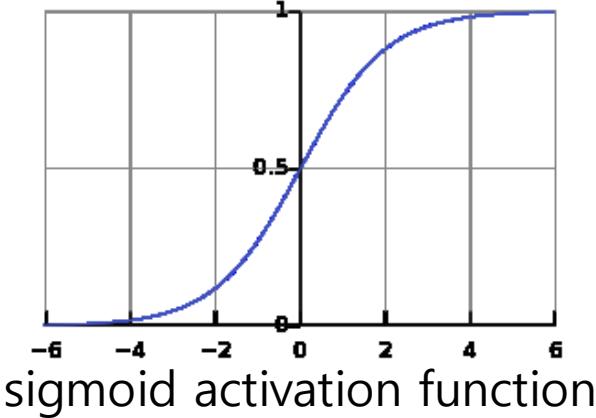
Perceptron(Artificial Neural Network)



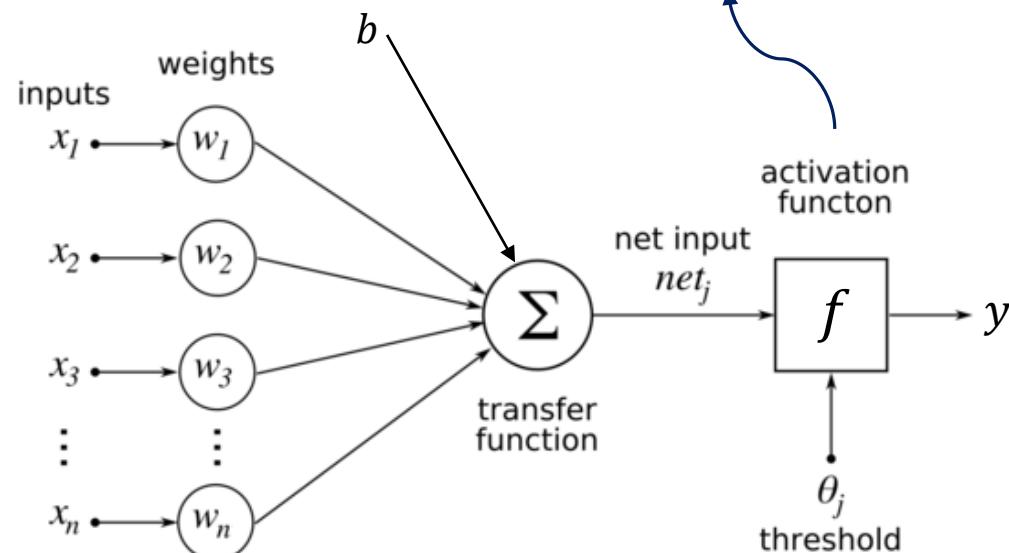
$$y = f(\mathbf{w}\mathbf{x} + b)$$

$$\mathbf{w} = [w_1 \ w_2 \ w_3 \ \dots \ w_n]$$

$$\mathbf{x} = [x_1 \ x_2 \ x_3 \ \dots \ x_n]^T$$

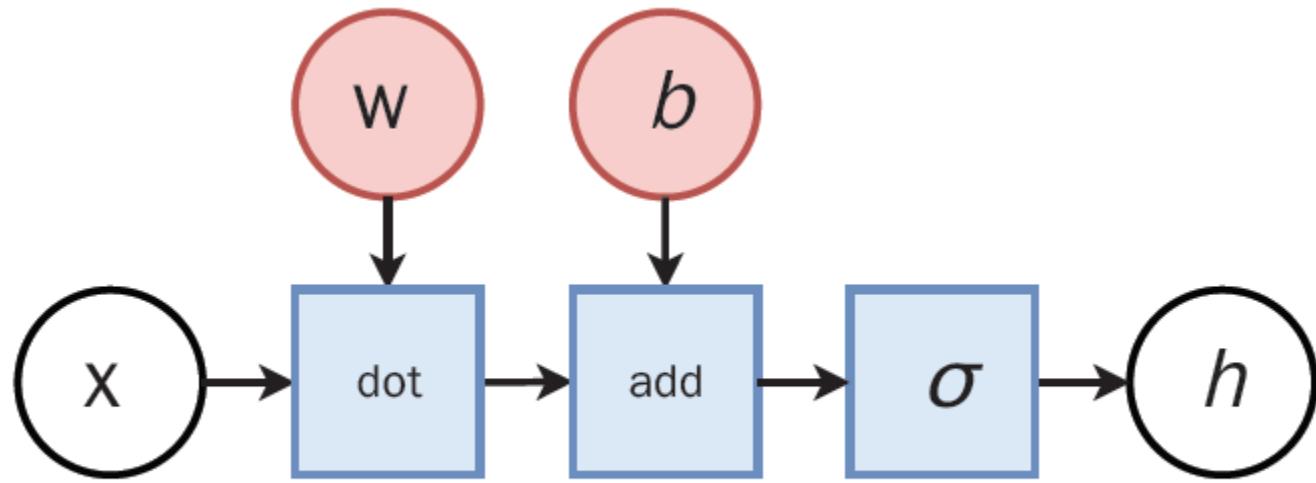


$$f(x) = \frac{1}{1 + e^{-x}}$$



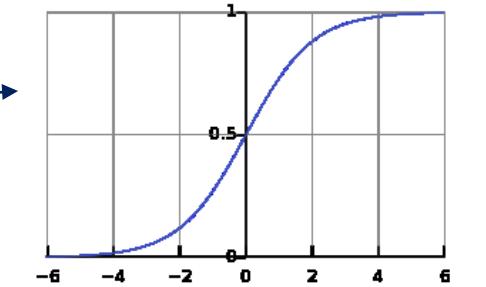
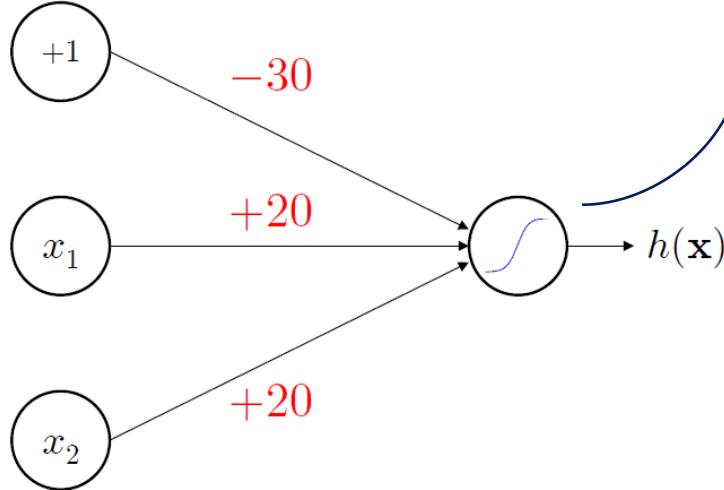
<Perceptron>

Computation Graph



This unit is the **lego brick** of all neural networks!

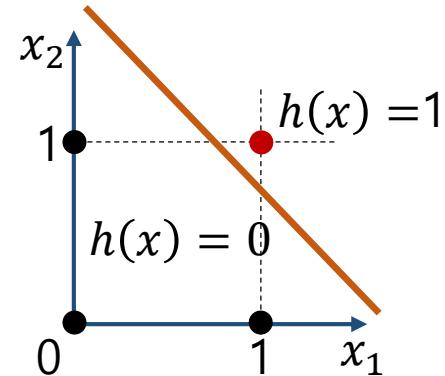
Example of ANN(logical AND)



| x_1 | x_2 | $h(\mathbf{x})$ |
|-------|-------|-------------------------|
| 0 | 0 | $\sigma(-30) \approx 0$ |
| 0 | 1 | $\sigma(-10) \approx 0$ |
| 1 | 0 | $\sigma(-10) \approx 0$ |
| 1 | 1 | $\sigma(10) \approx 1$ |

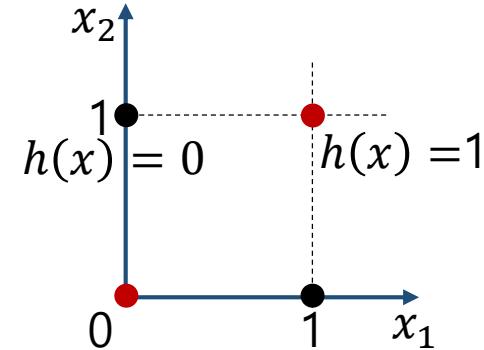
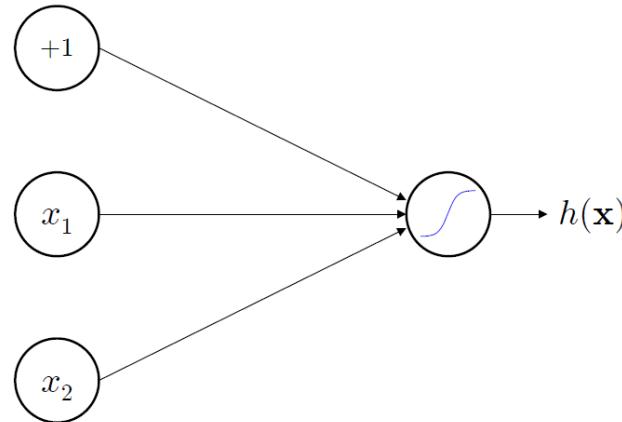
$$h(\mathbf{x}) = \sigma(-30 + 20x_1 + 20x_2)$$

학습이란 이러한 weight 값(-30, 20, 20)을
기계 스스로 찾을 수 있도록 해주는 과정!

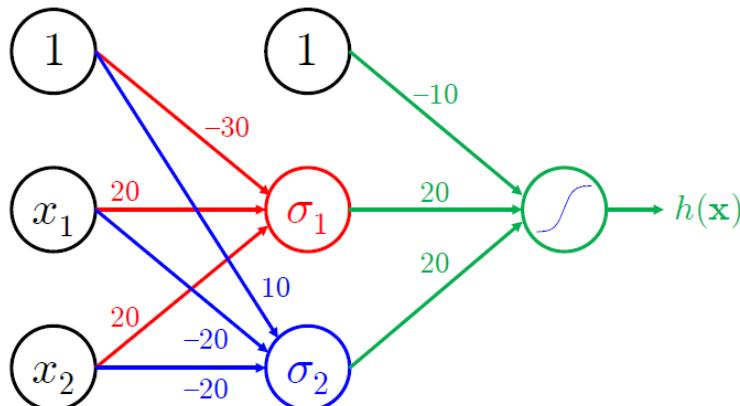


Logical XNOR with Perceptron

- Is it possible(Linearly separable)?



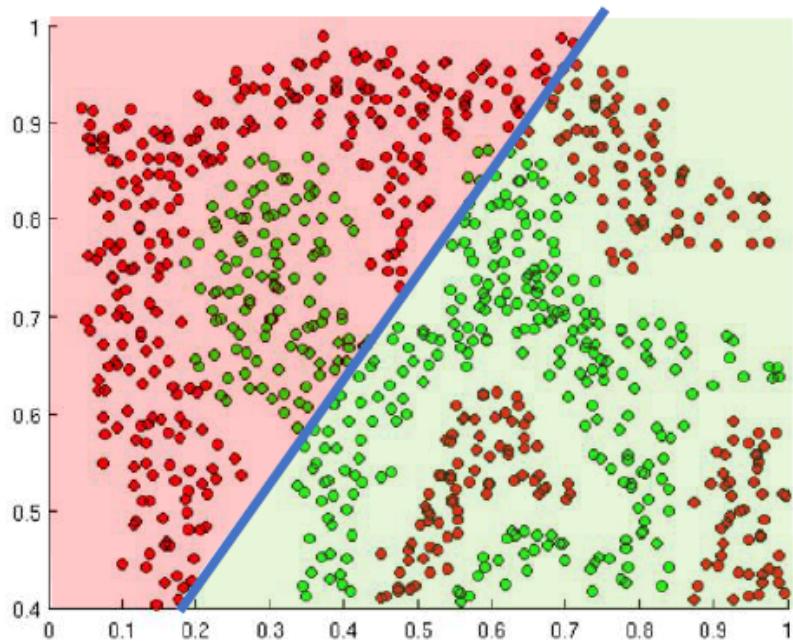
- We need more perceptrons and more layers



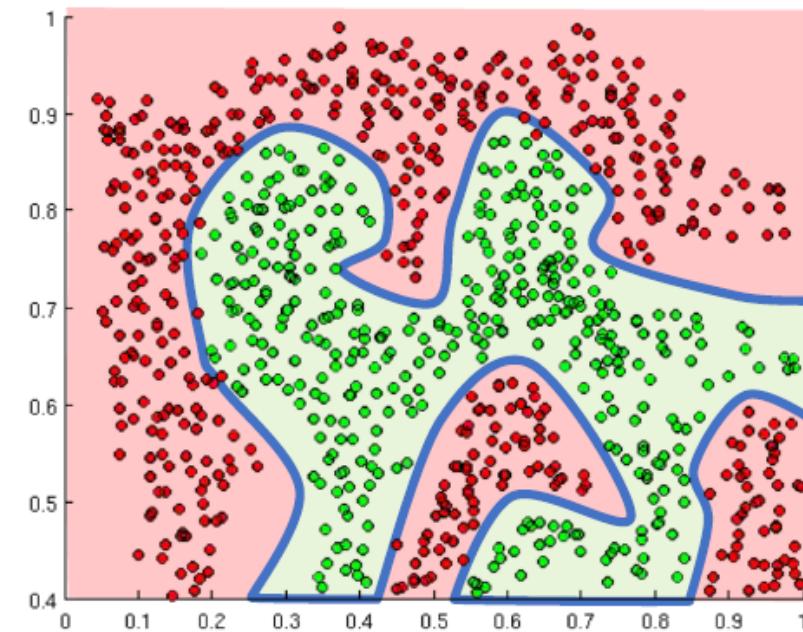
| x_1 | x_2 | σ_1 | σ_2 | $h(\mathbf{x})$ |
|-------|-------|------------|------------|-----------------|
| 0 | 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 |

Importance of Activation Functions

The purpose of activation functions is to *introduce non-linearities* into the network



Linear Activation functions produce linear decisions no matter the network size



Non-linearities allow us to approximate arbitrarily complex functions

아파트 가격 예측(Regression)

- 아파트 가격을 예측하는 프로그램을 작성해보자
 - 평수(평)
 - 층수(층)
 - 가장 가까운 지하철 역까지의 거리(km)
 - 해당 지역의 1인당 소득 평균(천만원)



| 평수 | 층수 | 지하철 역까지의 거리 (km) | 해당지역 1인당 소득 (천만원) | 아파트가격 (천만원) |
|----|----|------------------------|-------------------------|----------------|
| 34 | 15 | 1 | 4 | 72 |
| 32 | 4 | 5 | 3.3 | 60 |
| 18 | 9 | 2 | 2.5 | 34 |
| 42 | 3 | 2.5 | 5 | 100 |
| 21 | 10 | 1.5 | 3 | 42 |

일반적인 방법

- 기존의 data와 경험을 바탕으로 한 정보
 - 평수는 클수록 비싸다
 - 층수는 높은 층이 낮은 층보다 비싸다
 - 지하철 역까지의 거리는 가까울수록 비싸다
 - 그 지역의 소득수준이 높을수록 비싸다
- 위 정보들 간의 중요도에 따라 가중치를 주어서 최종 가격을 추론 한다
 - 평수가 가장 중요함
 - 층 수는 평수에 비해서 덜 중요함
 - 지하철 역까지의 거리는 중요하지만 거리값이 작을수록 가격이 비싸짐
 - 그 지역의 소득수준은 아파트 가격 결정에 가장 덜 중요함 등
- 예) 평수 $\times 2$ + 층수 $\times 0.3$ + 지하철 역까지 거리 $\times (-1)$ + 소득수준 $\times 0.1 =$ 아파트 가격
(천만원)

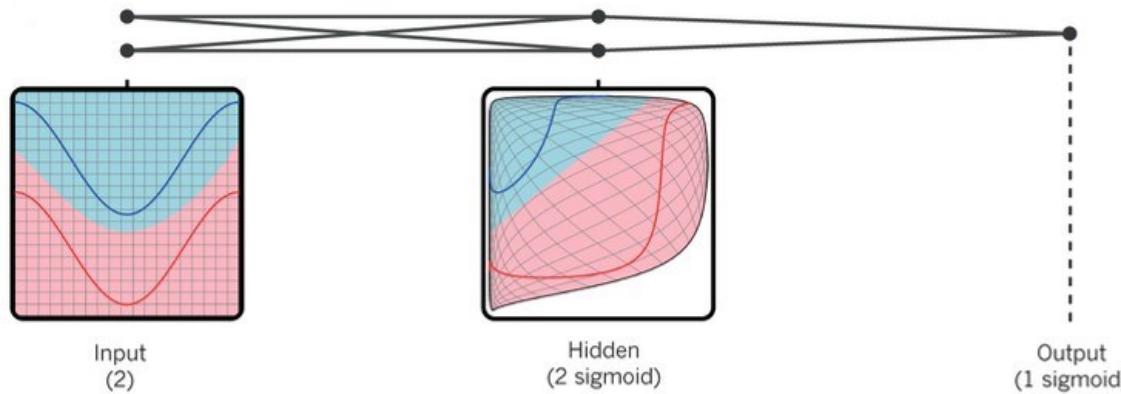
기계학습을 이용하는 방법

- 기존의 data를 이용하여 각 항목별로 □,△,○,☆에 들어갈 값을 기계 스스로 학습하게 한다.
 - 평수 × □ + 층수 × △ + 지하철 역까지 거리 × ○ + 소득수준 × ☆
= 아파트 가격(천만원)
- 어떻게?
 - 맨 처음 □,△,○,☆ 값을 random으로 정한 후 실제 가격과 비교하여 그 (차이값)²의 평균 계산한다
 - □,△,○,☆를 모두 1로 했을 경우, (실제가격 - 추론가격)²의 평균 = 287.524

| 평수 | 층수 | 지하철 역까지의 거리 (km) | 해당지역 1인당 소득 (천만원) | 아파트가격 (천만원) | 추론한 아파트 가격 (천만원) |
|----|----|------------------|-------------------|-------------|------------------|
| 34 | 15 | 1 | 4 | 72 | 54 |
| 32 | 4 | 5 | 3.3 | 60 | 44.3 |
| 18 | 9 | 2 | 2.5 | 34 | 31.5 |
| 42 | 3 | 2.5 | 5 | 100 | 52.5 |
| 21 | 10 | 1.5 | 3 | 42 | 35.5 |

- 이제 □,△,○,☆를 기계 스스로 조금씩 조정하여 (차이값)²의 평균이 0이 되게 할 수 있다면, 프로그램이 완성된다

Multi Layer Perceptron(MLP)



- 아파트 가격 결정에 소득수준이 높은 지역에서는 낮은 지역보다 지하철 역까지의 거리가 덜 중요하다면?
- 아파트 가격을 우선 여러가지로 예측하고 그 예측한 값들을 다시 잘 조합하여 더 잘 예측해보자(regression)
- 선을 잘 긋고 input 공간을 잘 왜곡하고 합하는 과정을 반복해서 데이터들을 잘 구분해보자(classification)
- 이렇게 perceptron을 여러 층으로 쌓으면 더 복잡한 문제를 풀 수 있다
- Linear fitting과 Non-linear transform의 반복

Universal Function Approximation

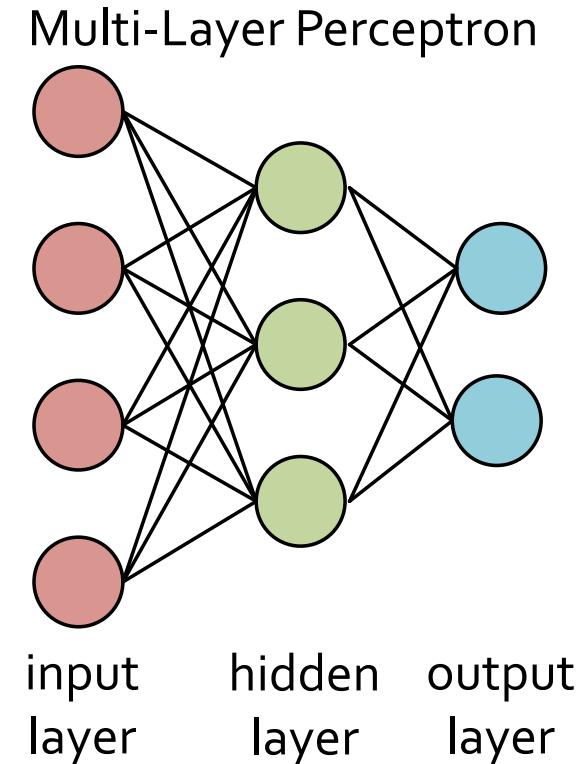
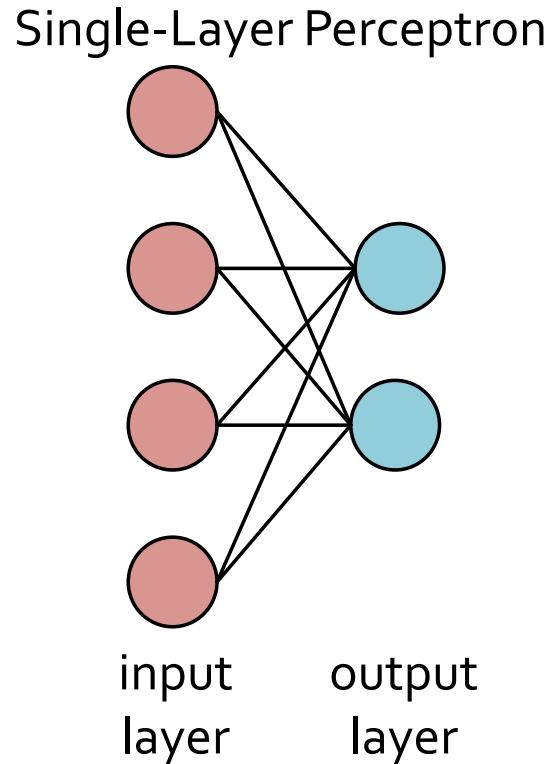
A feedforward network with a single layer is sufficient to approximate, to an arbitrary precision, any continuous function.

Caveats:

The number of hidden units may be infeasibly large

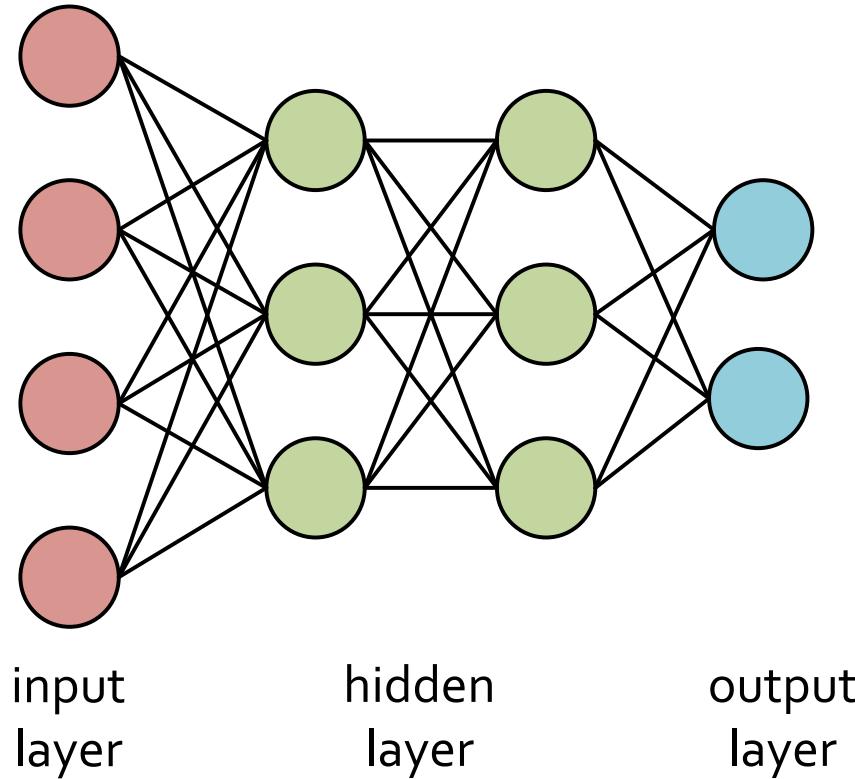
The resulting model may not generalize

Single-Layer Perceptron & Multi-Layer Perceptron(MLP)

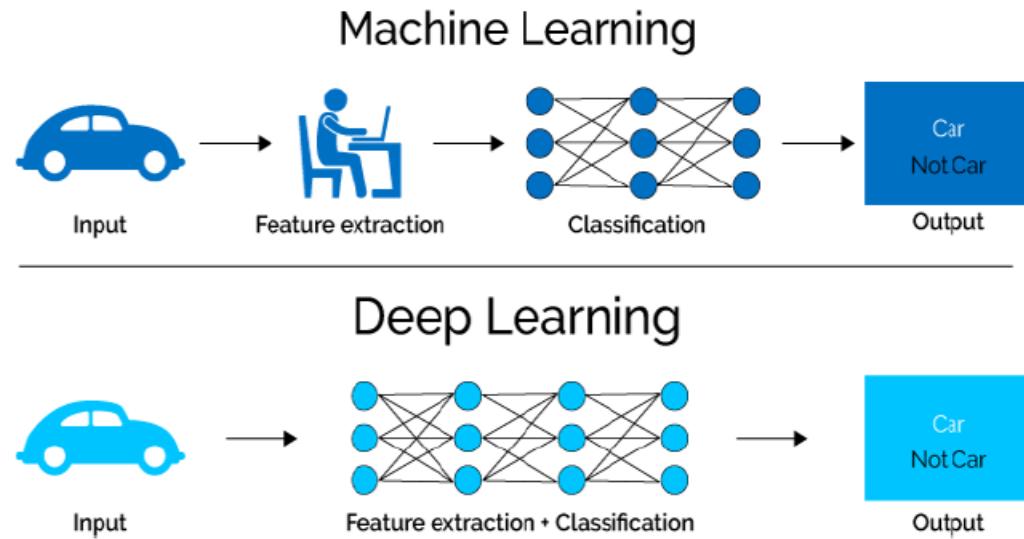


Deep Learning

- Deep Neural Network을 이용한 Machine Learning 방법
 - Deep Neural Network : Hidden layer 수가 2개 이상인 network



Machine Learning vs Deep Learning



ML : Audio → Acoustic Model → Phonetic Model → Language Model → Text

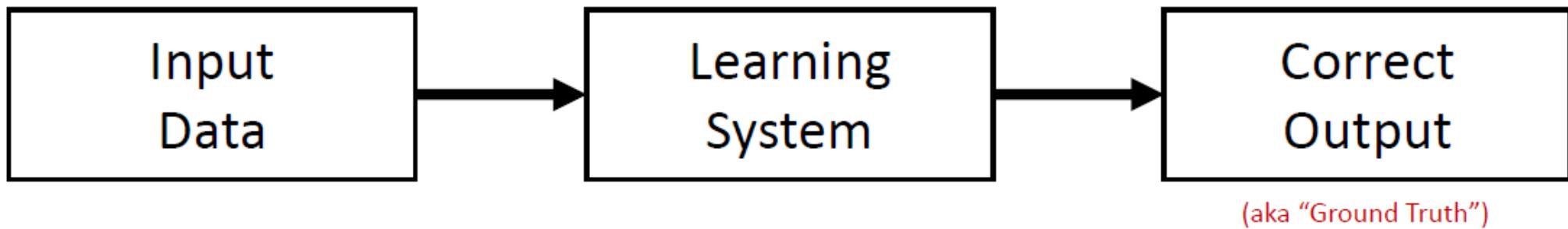
DL : Audio → Deep Networks → Text (End to End Learning)

ML : Pixels → Key Points → SIFT Features → Deformable Part Model → Label

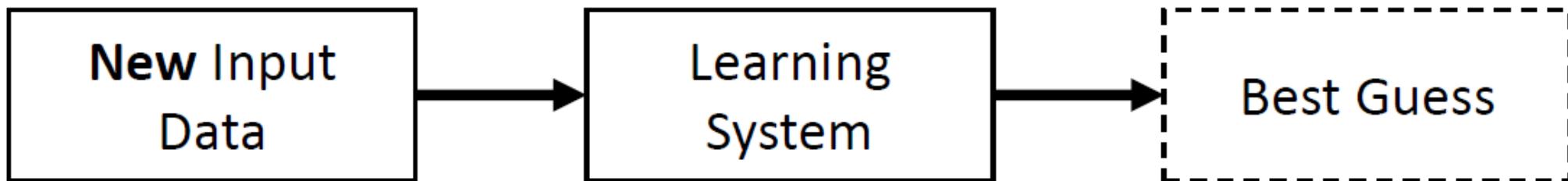
DL : Pixels → Deep Networks → Label (End to End Learning)

Training and Testing

Training Stage:



Testing Stage:

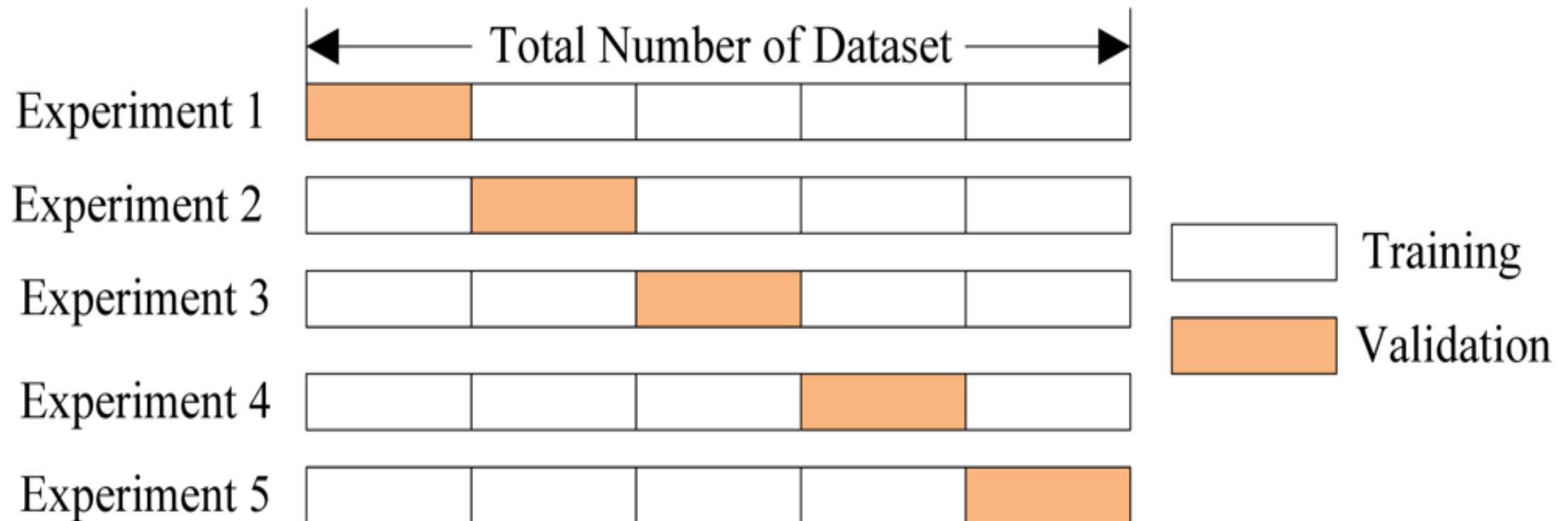


Dataset

- Training set, Test set
 - or
- Training set, Validation set, Test set
- Training → Tuning(validation) → Test
- In any case, DO NOT USE the test set before the training is over!

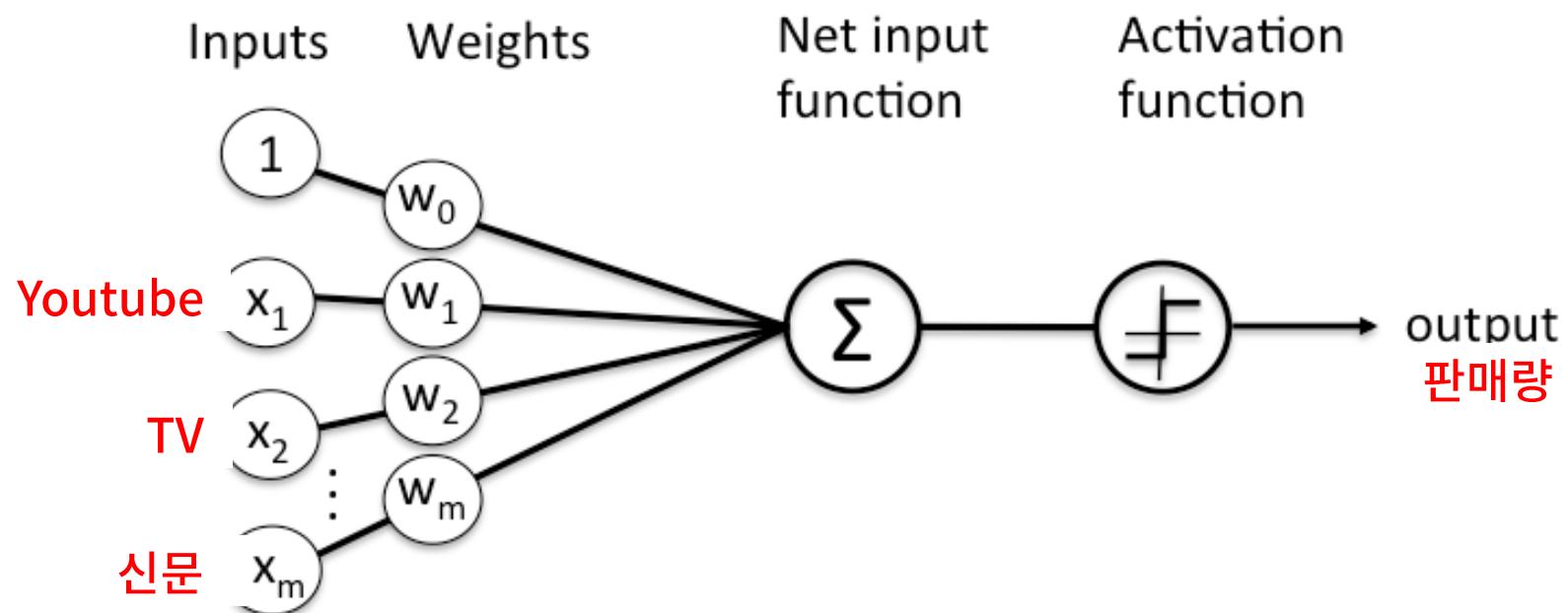


Cross-Validation



An Example of Perceptron

어떤 상품을 판매하는 회사에서
Youtube, TV, 신문 광고료에 따른 판매량을 예측



| Youtube | TV | 신문 | 판매량 |
|---------|-----|-----|-----|
| 130 | 70 | 55 | 221 |
| 44 | 40 | 45 | 109 |
| 17 | 46 | 70 | 73 |
| 81 | 40 | 57 | 181 |
| 52 | 12 | 11 | 90 |
| 9 | 49 | 61 | 52 |
| 57 | 33 | 23 | 122 |
| 30 | 30 | 25 | 100 |
| ... | ... | ... | ... |

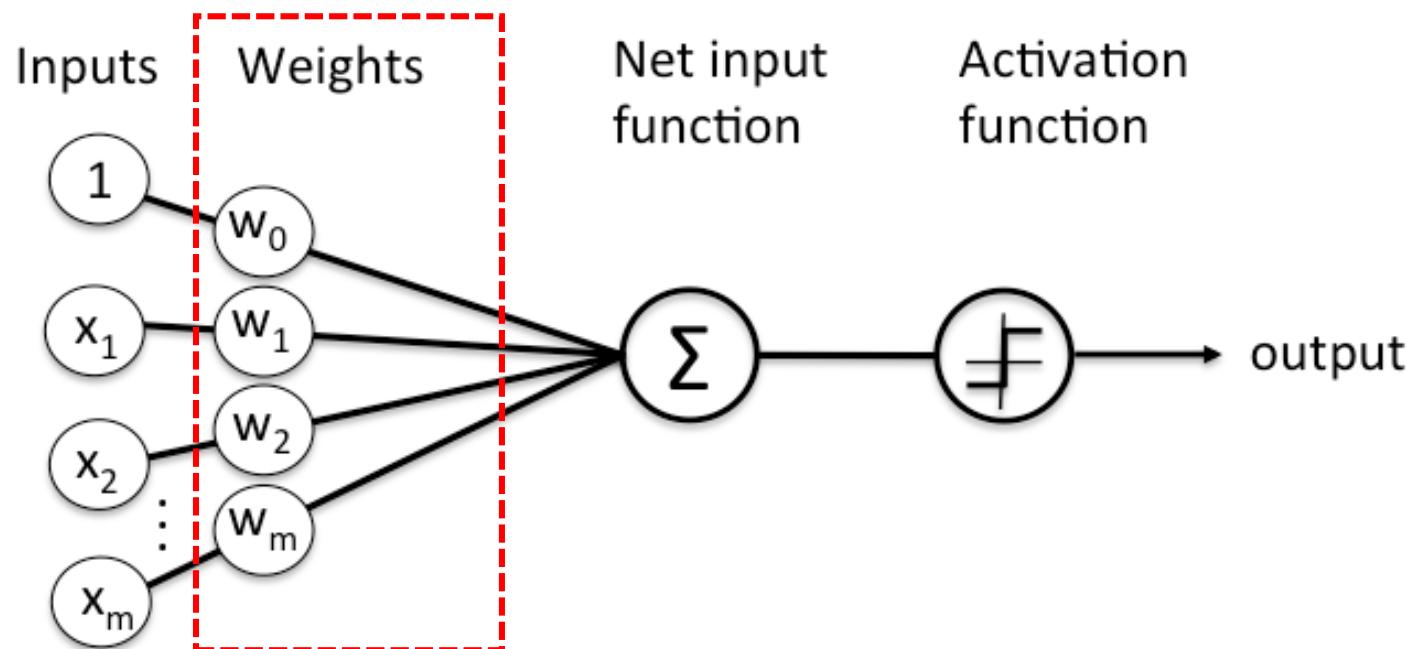
Weight Initialization

원하는 Weight를 어떻게 찾을까?

그 전에 아무것도 모르는 상태에서 Weight의 초기값은 어떻게?

잘 모르겠으니 일단 Random한 값으로 시작하자!

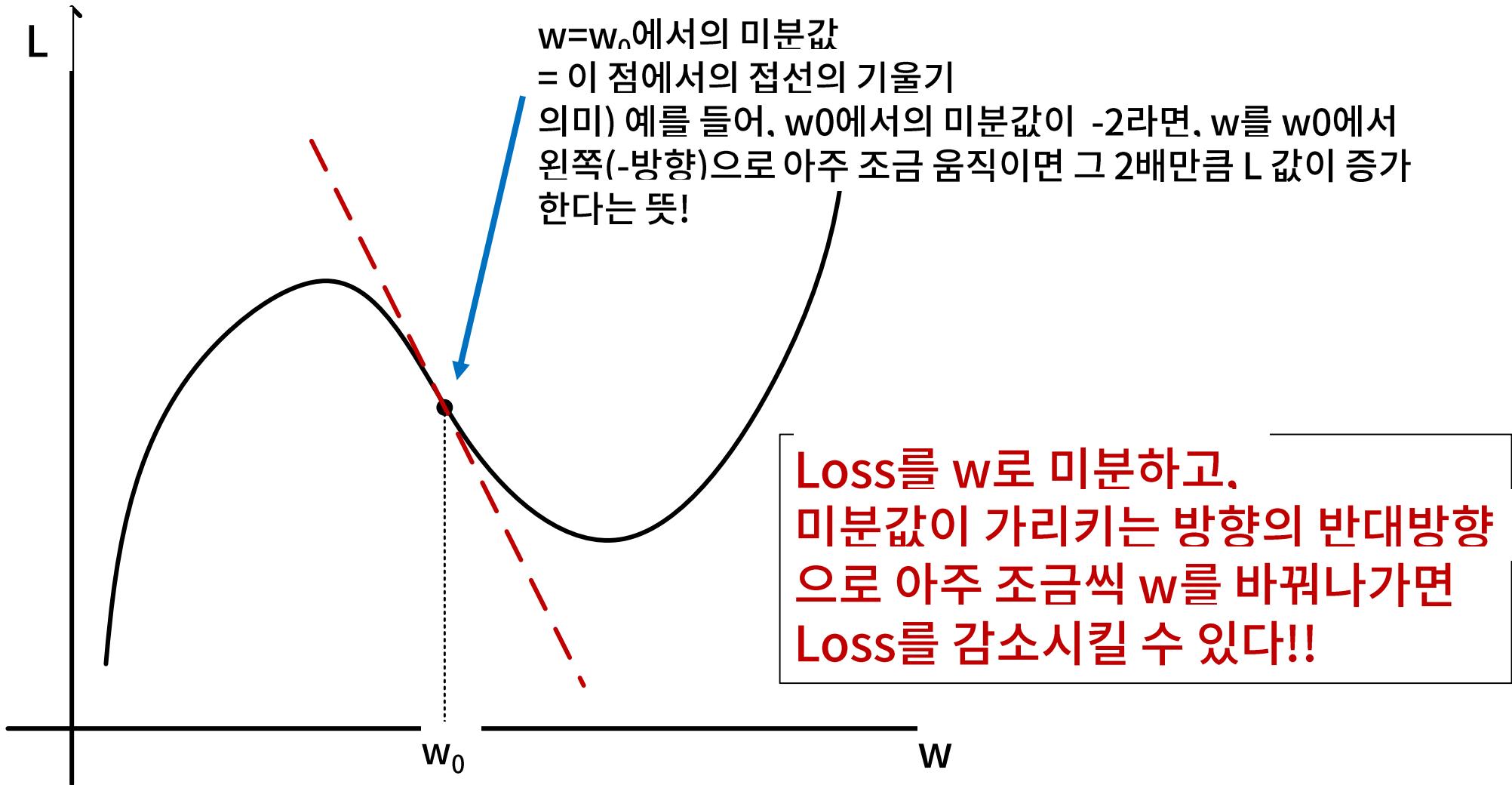
– Random Initialization!



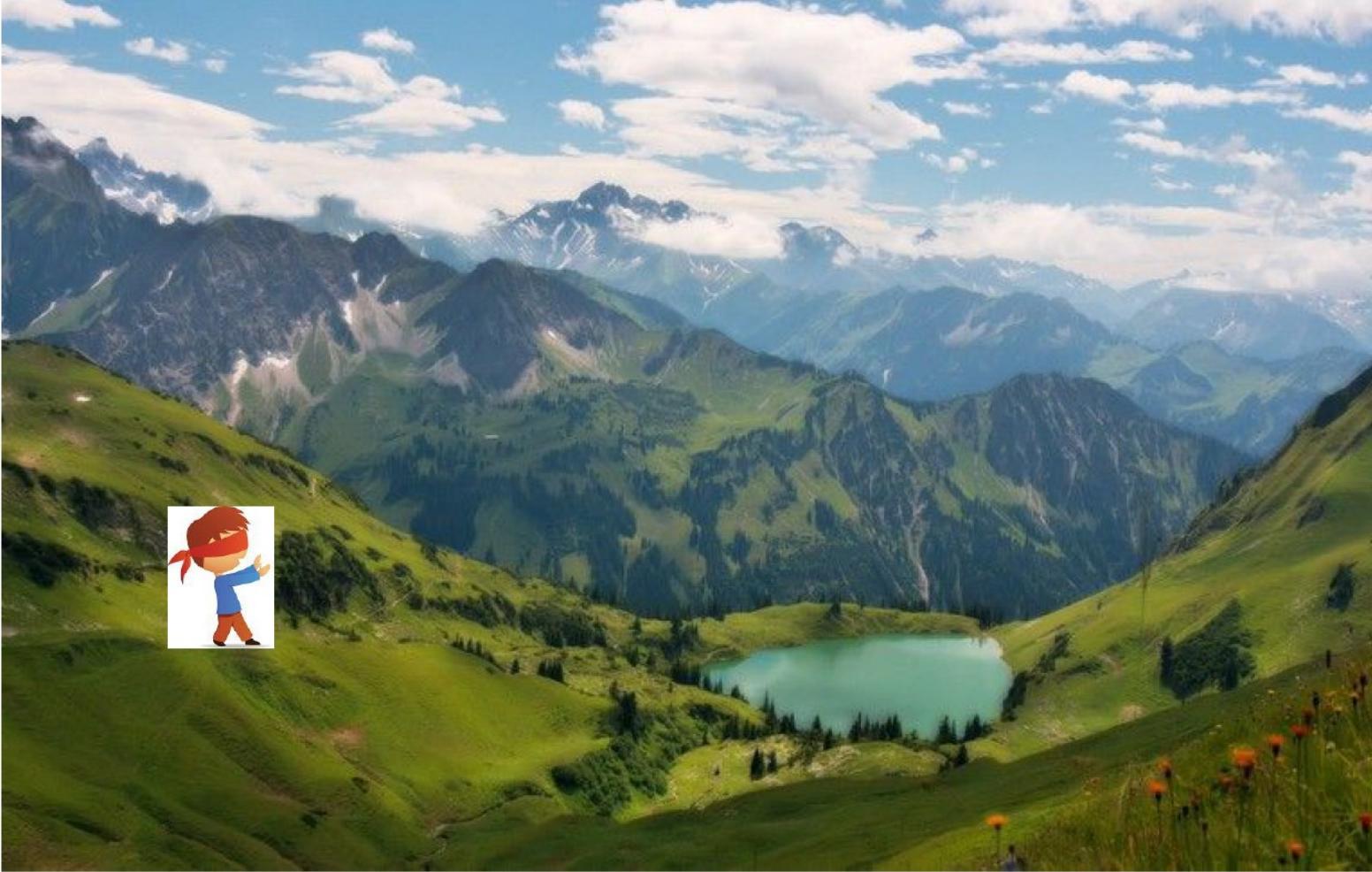
Loss Function(Cost Function)

- Neural Network이 얼마나 잘하는 지 or 얼마나 못하는 지에 대한 척도가 필요함
 - Loss Function / Cost Function
- 많이 쓰는 방법 중 한가지
 - (Neural Network의 출력과 실제 정답의 차이)²
 - 예) 2가지 광고료 조합에 대하여 Neural Network이 예측한 판매량 : (100, 80), 실제 판매량 : (105, 78)일 때,
$$L = (105 - 100)^2 + (78 - 80)^2 = 29$$
- 이제 필요한 것은 Loss Function의 값이 줄어들도록 weight 값을 조금씩 바꾸는 것
 - Weight를 어떻게 바꿔야 Loss Function 값이 줄어들까? → 미분!

미분??



Gradient Descent



Gradient Descent

Loss Function의 미분(Gradient)를 이용하여 weight를 update하는 방법

$$w_{new} = w_{old} - \eta \nabla_w L$$

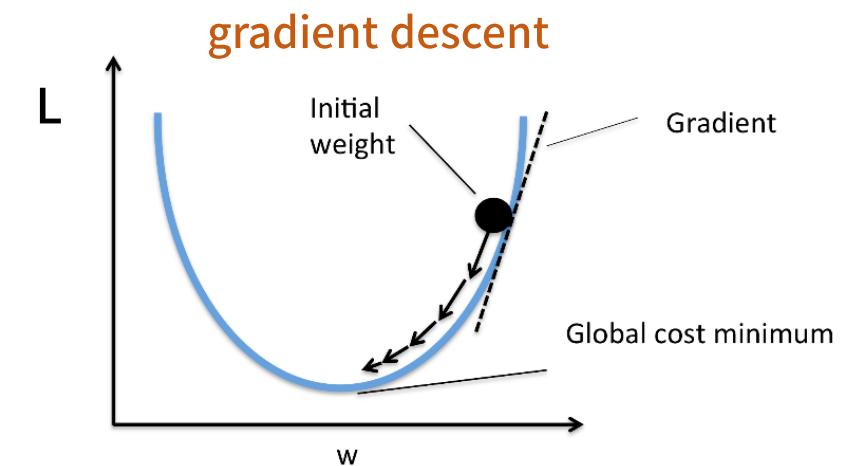
Weight update = $w_{new} - w_{old}$

$$= -\eta \nabla_w L$$

Loss를 감소 시키는 방향 (Descent)

아주 조금씩 이동 (Learning Rate)

미분값 (Gradient)



Gradient Descent

- 내가 가진 모든 training data에 대하여 Neural Network의 출력과 실제 정답(label)을 비교하여 각각의 Loss를 계산하고 이를 모두 더해서 전체 Loss를 계산한다
- 그리고, 이 loss를 weight로 미분한 다음 그 미분값(gradient)가 가리키는 방향의 반대방향으로 weight 값을 아주 조금씩 바꿔나간다
- 그런데 내가 가진 data가 아주아주 많은데 weight 값을 아주 조금 바꾸기 위해서 모든 data에 대해서 loss를 다 계산해야 한다...
- Data가 더 많아질수록 학습이 더 느려진다...

Batch / Stochastic / Mini-batch GD

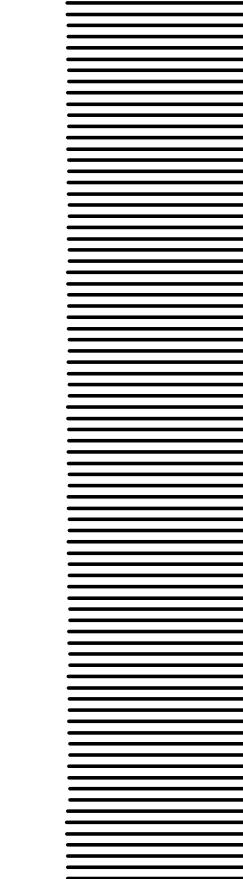
모든 training data에 대하여 loss를 다 계산하는 건 너무 오래 걸리니까
전체 data 중에 일부 data만 뽑고(sampling) 이 data들이 전체 training
data를 잘 대표한다고 가정해서 이 data들로만 loss를 계산하여 weight
를 update해도 되지 않을까?

- Batch Gradient Descent
 - 모든 data에 대해서 loss를 계산하여 다 더하고 이를 이용해서 GD
- Stochastic Gradient Descent
 - Data를 1개만 뽑고, 그 1개의 data에 대한 loss를 이용하여 GD
- Mini-batch Gradient Descent
 - Batch/Stochastic의 중간 형태로 data를 n개를 뽑고 그 n개의 data에 대한 loss
를 계산하여 다 더한 뒤 이를 이용하여 GD → 제일 많이 쓰는 방법!

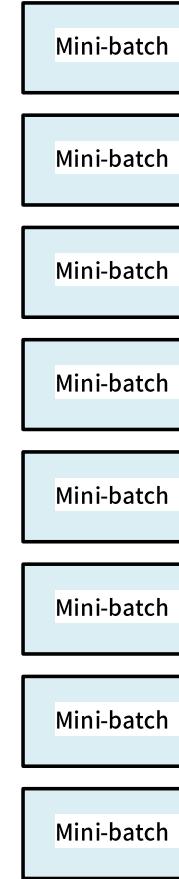
Batch / Stochastic / Mini-batch GD



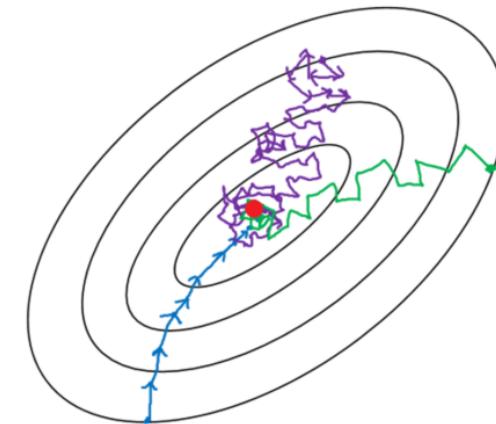
Batch
Gradient Descent



Stochastic
Gradient Descent



Mini-batch
Gradient Descent



- Batch gradient descent
- Mini-batch gradient Descent
- Stochastic gradient descent

Gradient Descent Again

Loss Function의 미분(Gradient)를 이용하여 weight를 update하는 방법

$$w_{new} = w_{old} - \eta \nabla_w L$$

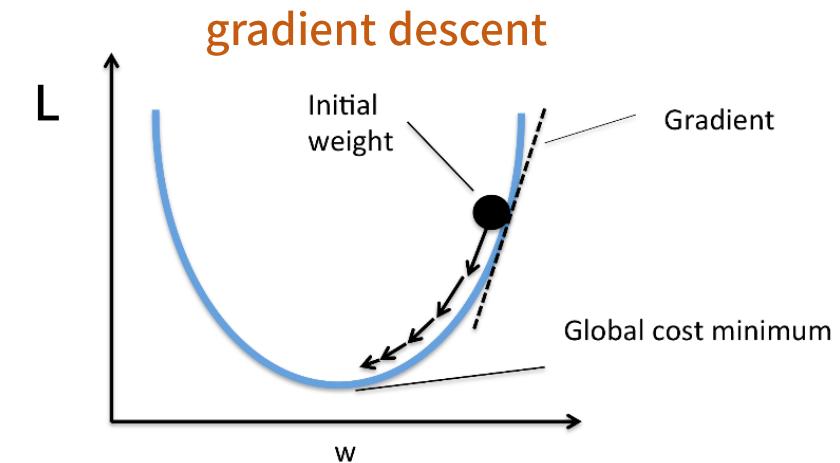
Weight update = $w_{new} - w_{old}$

$$= -\eta \nabla_w L$$

Loss를 감소 시키는 방향
(Descent)

아주 조금씩 이동
(Learning Rate)

미분값
(Gradient)



미분을 계산해봅시다!

$$z_{11} = x_1 \cdot w_{11} + x_2 \cdot w_{12} + x_3 \cdot w_{13} + x_4 \cdot w_{14}$$

$$a_{11} = \sigma(z_{11}) = \frac{1}{1 + e^{-z_{11}}}$$

$$z_2 = a_{11} \cdot w_{21} + a_{12} \cdot w_{22} + a_{13} \cdot w_{23}$$

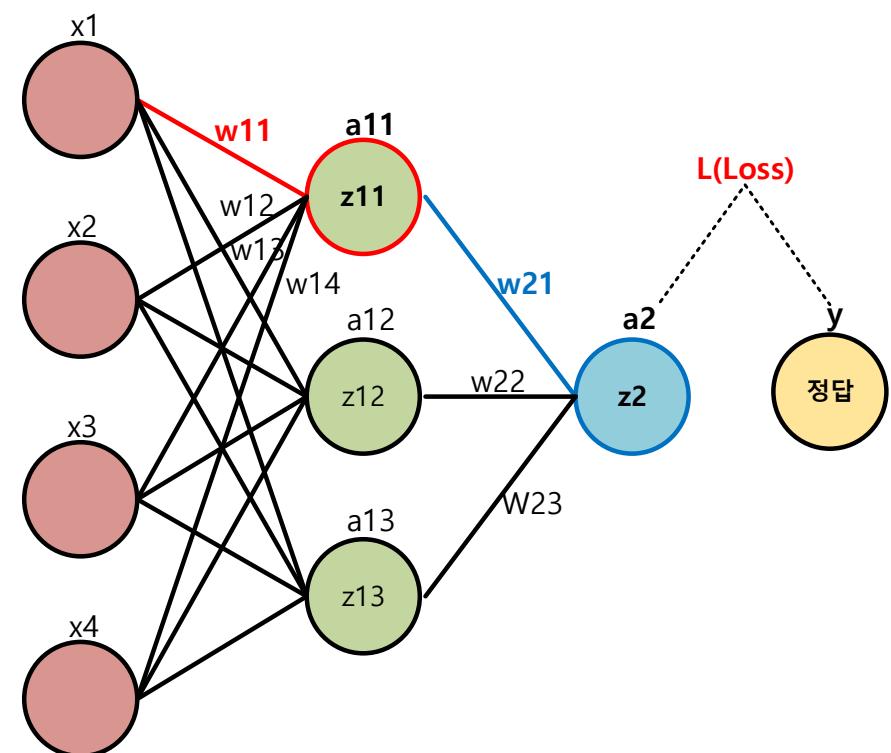
$$a_2 = z_2$$

$$L = (y - a_2)^2$$

일 때,

w₁₁을 update 하기 위해 필요한 미분값

$$\frac{\partial L}{\partial w_{11}} = ??????$$



Back Propagation

$$z_{11} = x_1 \cdot w_{11} + x_2 \cdot w_{12} + x_3 \cdot w_{13} + x_4 \cdot w_{14}$$
$$a_{11} = \sigma(z_{11}) = \frac{1}{1 + e^{-z_{11}}}$$
$$z_2 = a_{11} \cdot w_{21} + a_{12} \cdot w_{22} + a_{13} \cdot w_{23}$$
$$a_2 = z_2$$
$$L = (y - a_2)^2$$

Loss부터 거꾸로 한 단계씩 미분을 해봅시다

$$\partial L / \partial a_2 = -2(y - a_2)$$

$$\partial a_2 / \partial z_2 = 1$$

$$\partial z_2 / \partial a_{11} = w_{21}$$

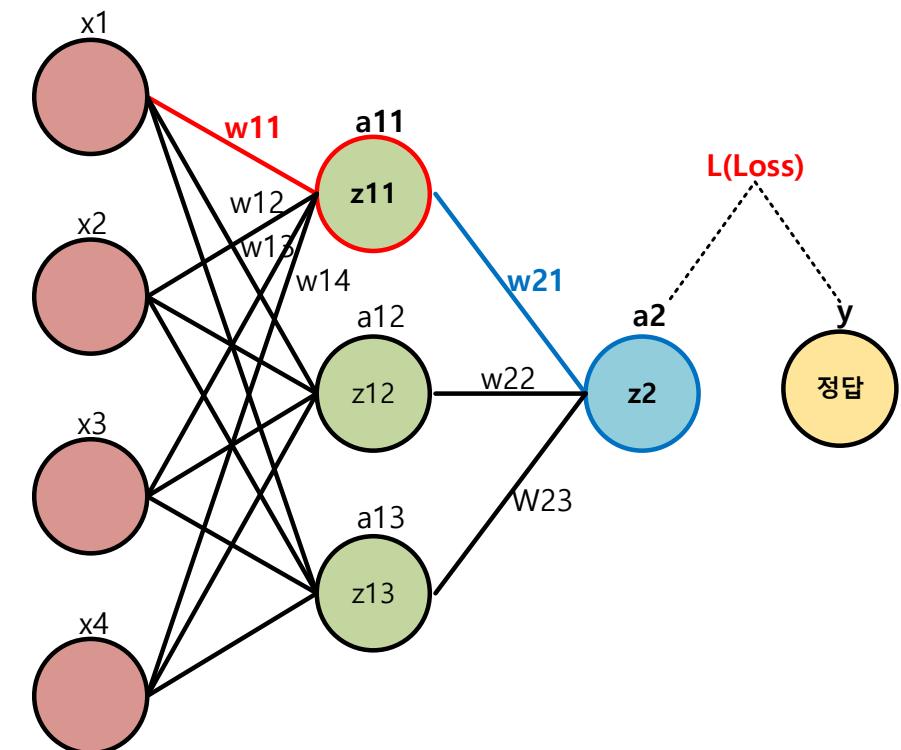
$$\partial a_{11} / \partial z_{11} = \sigma(z_{11}) \cdot (1 - \sigma(z_{11}))$$

$$\partial z_{11} / \partial w_{11} = x_1$$

이 미분들을 전부 각각 곱하면(chain rule),

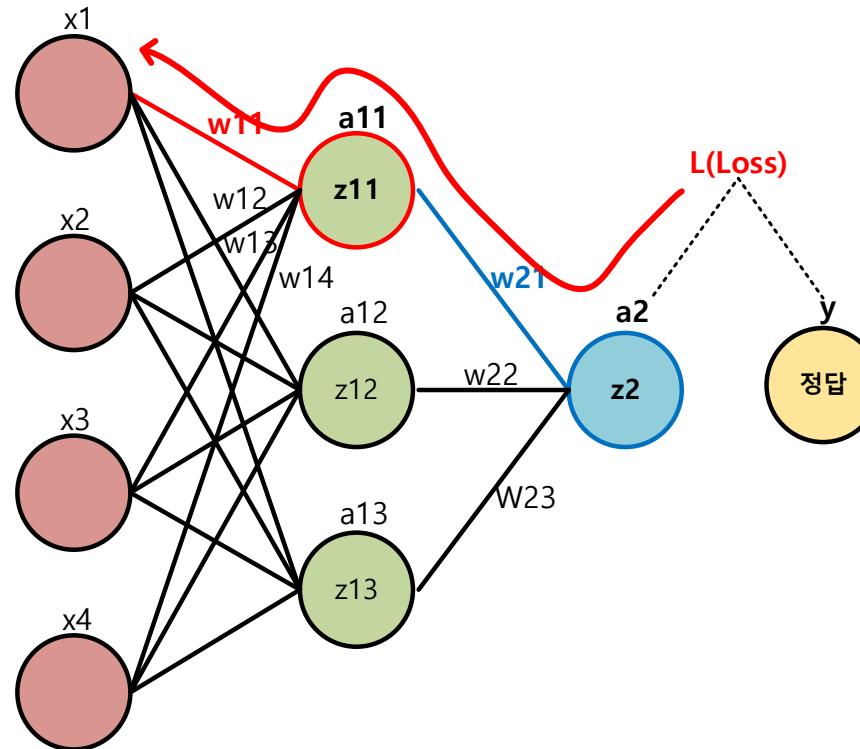
$$\frac{\partial L}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_{11}} \cdot \frac{\partial a_{11}}{\partial z_{11}} \cdot \frac{\partial z_{11}}{\partial w_{11}} = \frac{\partial L}{\partial w_{11}}$$

우리가 구하려고 했던 미분값



Back Propagation

Loss로부터 거꾸로 한 단계씩 미분 값을 구하고 이 값들을 chain rule에 의하여 곱해가면서 weight에 대한 gradient를 구하는 방법

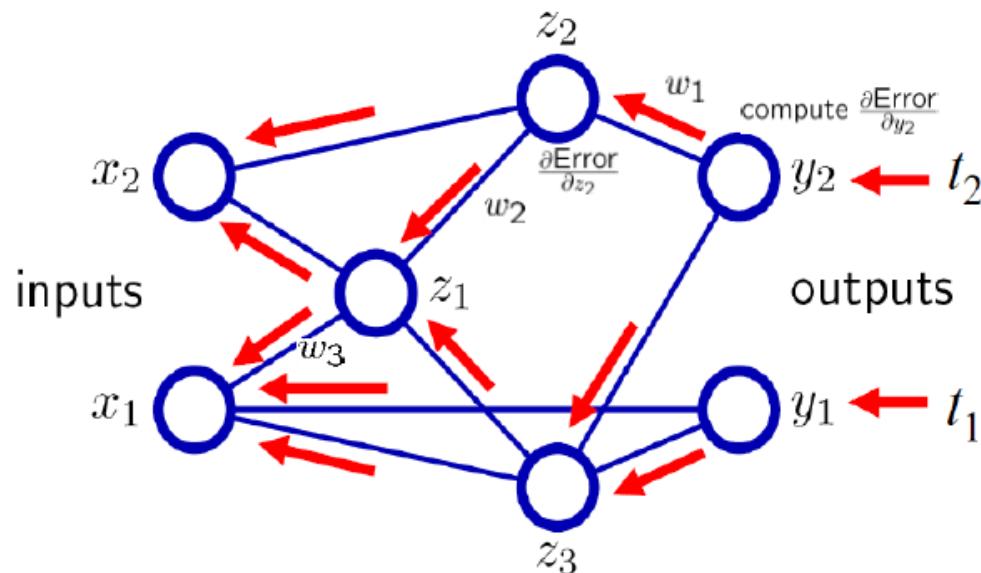
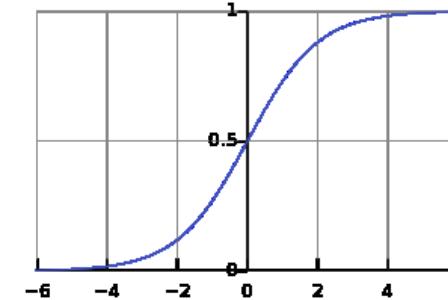


딥러닝을 어렵게 하는 것들

- Vanishing gradient problem
- Overfitting problem
- Get stuck in local minima

Vanishing Gradient Problem

- Gradient 값이 뒤로 전달될 수록 점점 작아짐
- Sigmoid 사용으로 인하여(미분값의 최대 : $\frac{1}{4}$) 아래쪽 layer는 학습이 이루어지지 않음



$$\frac{\partial \text{Error}}{\partial w_1} = \frac{\partial \text{Error}}{\partial y_2} \frac{\partial y_2}{\partial w_1} = \frac{\partial \text{Error}}{\partial y_2} \sigma'(\cdot) z_2$$

$$\frac{\partial y_2}{\partial w_1} = \sigma'(\cdot) z_2$$

$$\frac{\partial \text{Error}}{\partial w_2} = \frac{\partial \text{Error}}{\partial z_2} \frac{\partial z_2}{\partial w_2} = \frac{\partial \text{Error}}{\partial y_2} \sigma'(\cdot) w_1 \sigma'(\cdot) z_1$$

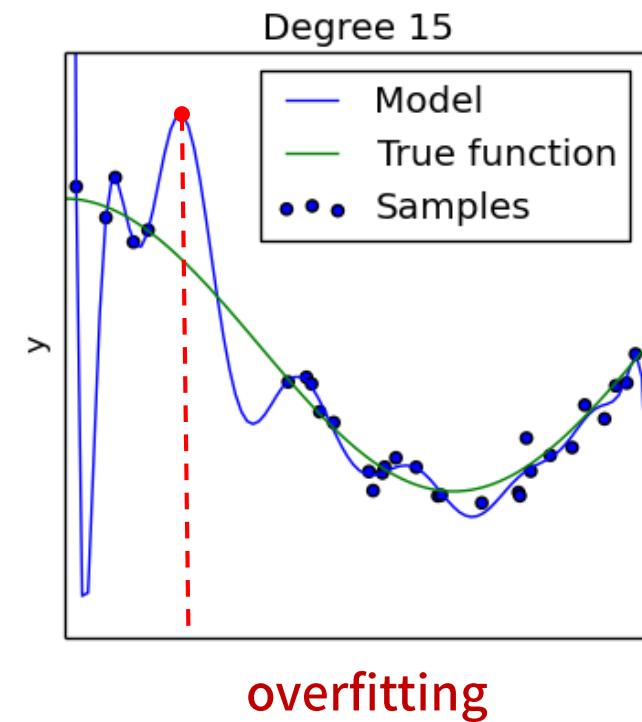
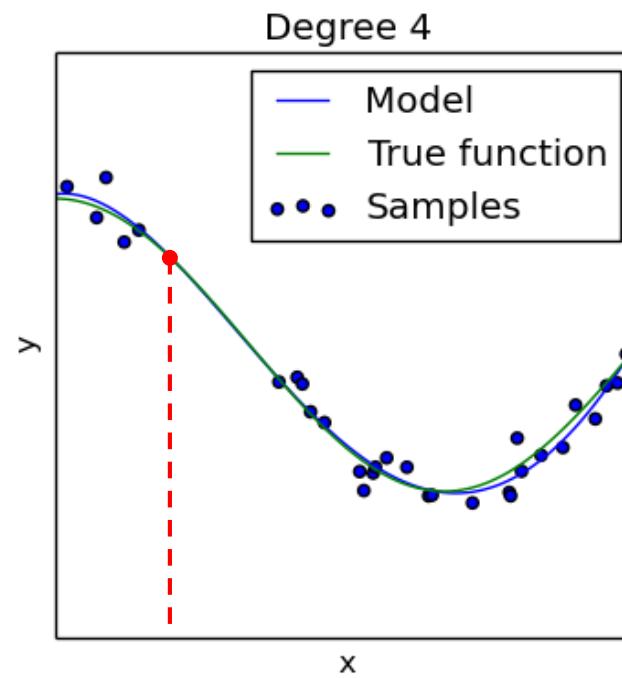
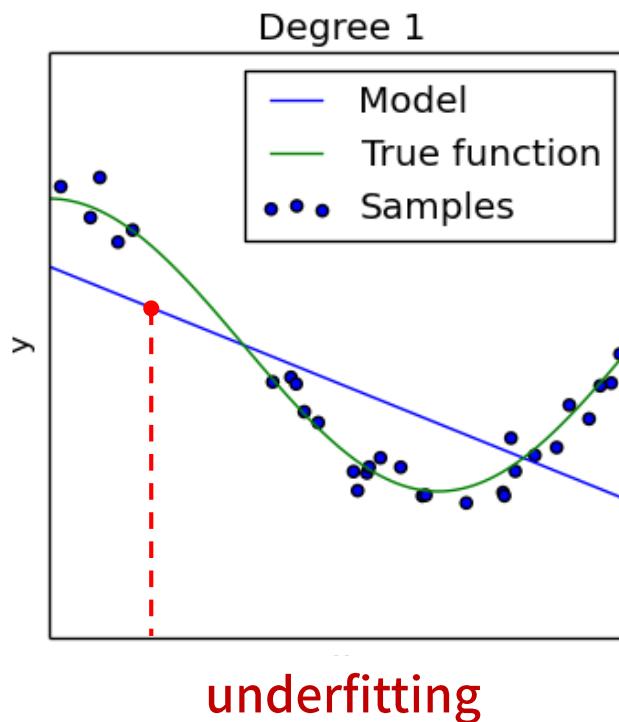
$$\frac{\partial z_2}{\partial w_2} = \sigma'(\cdot) z_1$$

$$\frac{\partial \text{Error}}{\partial w_3} = \sigma'(\cdot) \sigma'(\cdot) \sigma'(\cdot) C$$

$$\sigma'(\cdot) \approx 0$$

Overfitting Problem

- Data가 많지 않은 경우에 발생할 수 있음
- Overfitting : Training data에 너무 최적화(fitting)를 함으로 인해서 일반화(generalization) 성능이 떨어지는 현상



Overfitting Problem

열심히 Neural Network에게 고양이



를 가르쳤더니...



뚱뚱하니까 고양이 아님



갈색이니까 고양이 아님



귀 쳐졌으니까 고양이 아님

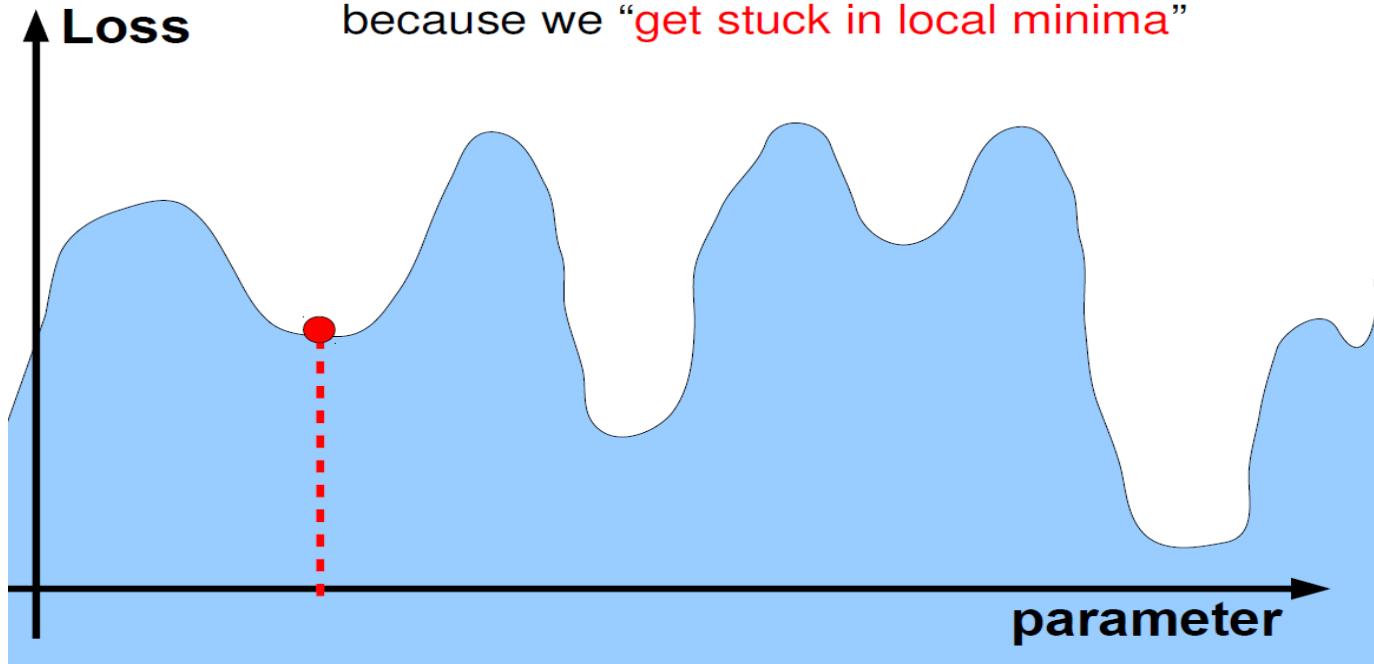
융통성이 하나도 없어짐 → Overfitting!!

Local Minima

- 어디서 시작하느냐에 따라서 잘못하면 local minima에 빠질 위험이 존재

ConvNets: till 2012

Common wisdom: training does not work because we “get stuck in local minima”

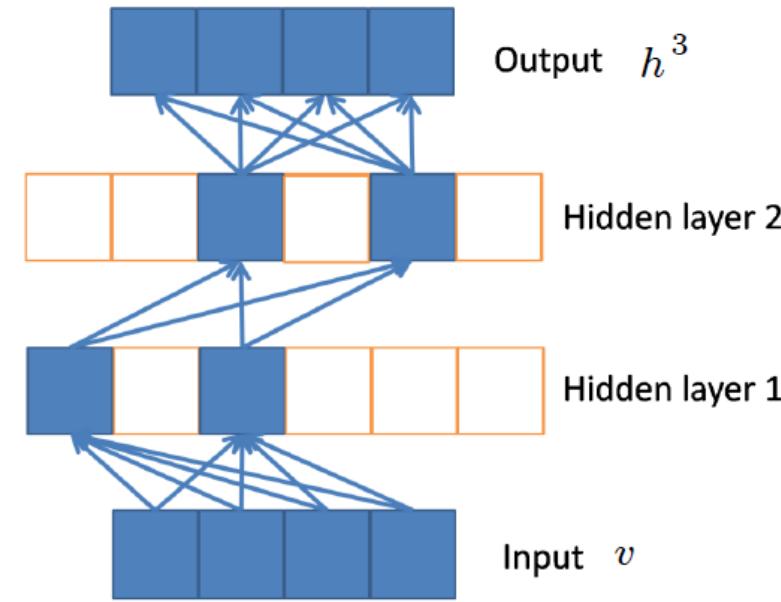
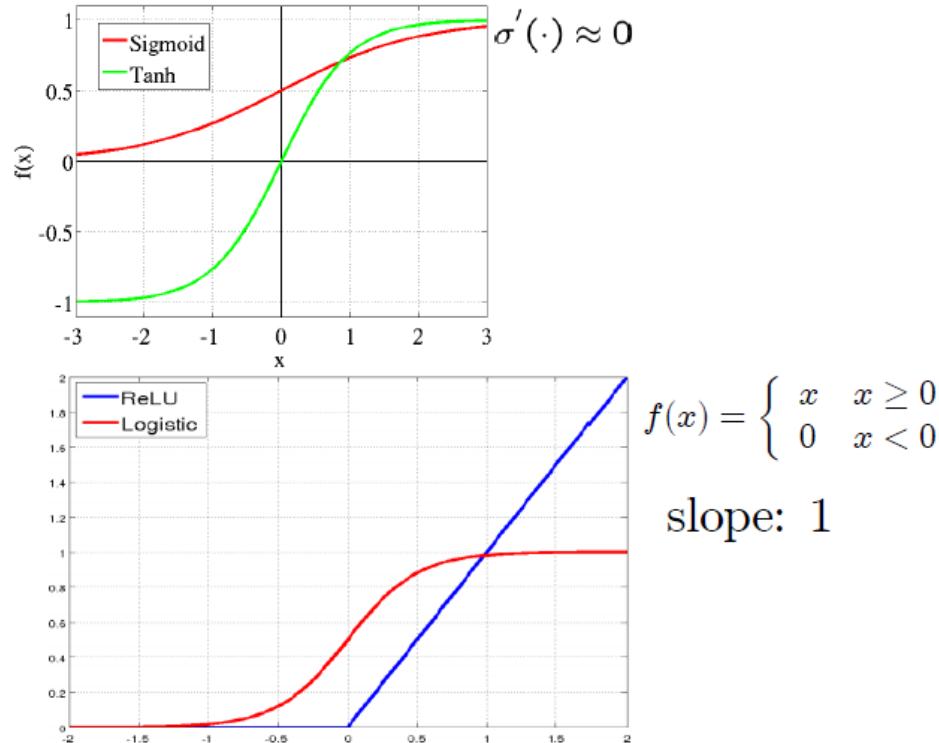


Solutions

- Vanishing gradient problem
→ Sigmoid 말고 ReLU를 쓰자
- Overfitting problem
→ Regularization method를 쓰자(예 : dropout)
- Get stuck in local minima
→ Local minima에 빠져도 괜찮다

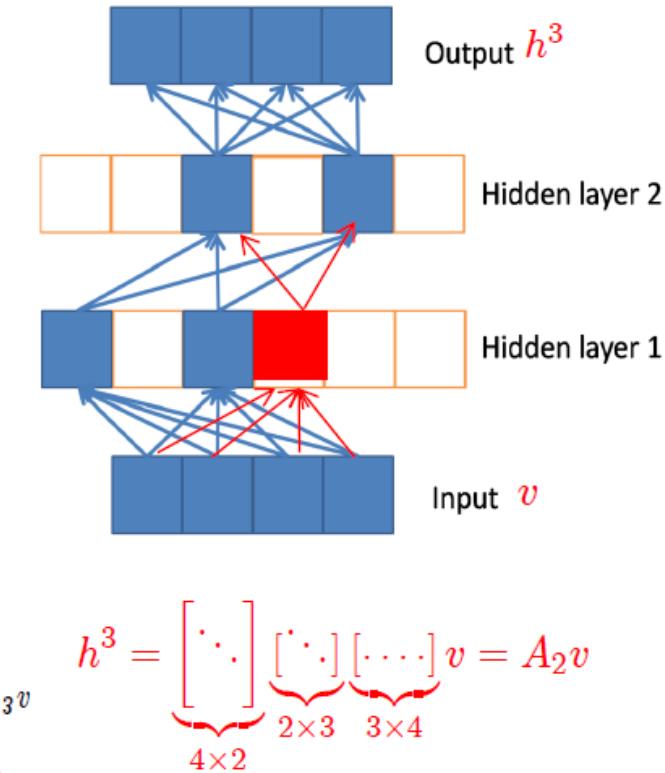
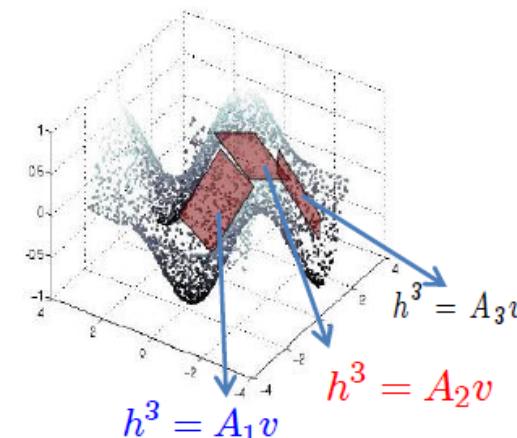
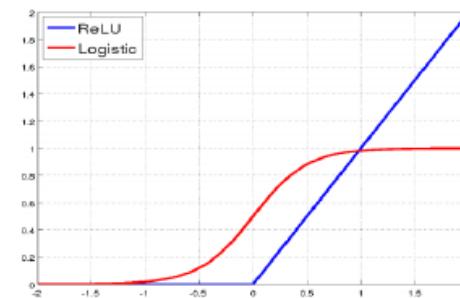
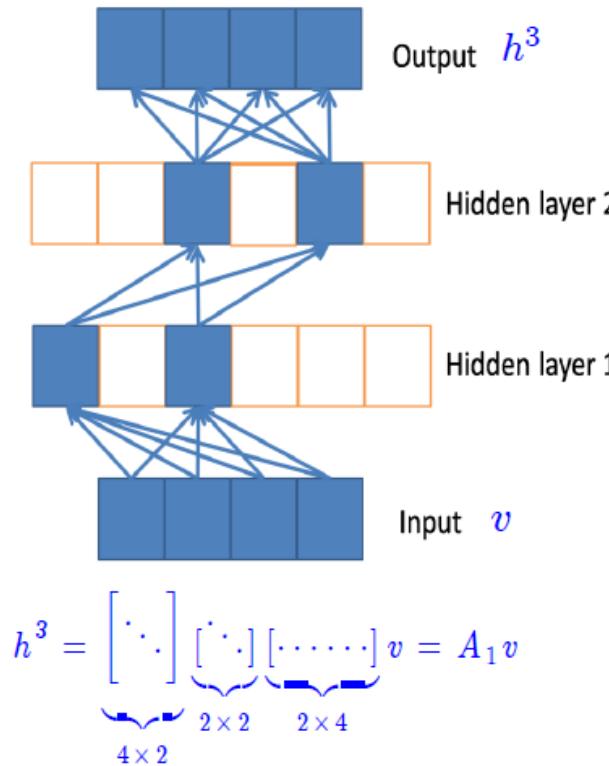
ReLU : Rectified Linear Unit

- ReLU를 activation function으로 사용 → sparse activation
- ReLU는 미분값이 0 아니면 1 → vanishing gradient 해결



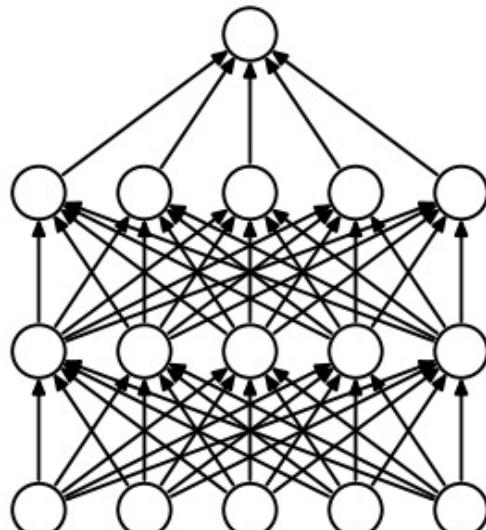
ReLU의 의미

- Piece-wise linear tiling : locally linear mapping

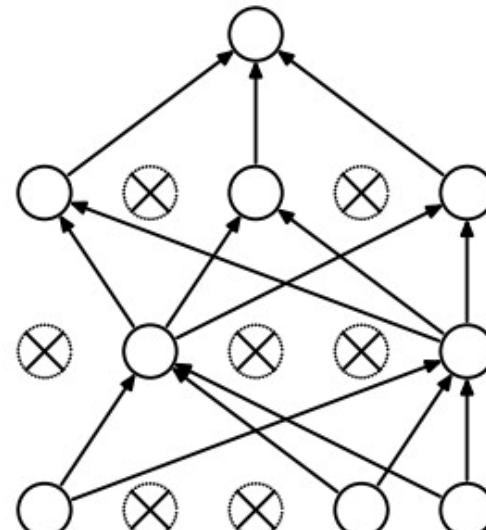


Dropout(Regularization Method)

- 각 학습 단계마다, 특정 확률로(예 : 50%) random하게 hidden layer에 있는 unit들을 없애고 학습하는 방법
- Ensemble 개념을 적용
 - 여러 개의 model을 사용하여 평균값을 쓰면 하나의 model을 쓰는 경우보다 좋음
 - 하나의 model로 비슷한 효과를 낼 수 있는 방법



(a) Standard Neural Net



(b) After applying dropout.

Other Regularization Methods

- Weight Decay(L2 Regularization)

$$E(w) = E_0(w) + \frac{1}{2} \lambda \sum_i w_i^2$$

- Batch Normalization

- Benefits of BN

- Increase learning rate
- Remove dropout
- Reduce L2 weight decay
- Remove LRN

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1 \dots m\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{mini-batch mean}$$

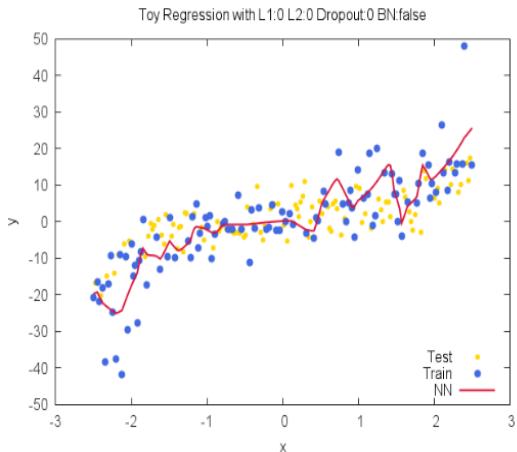
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{normalize}$$

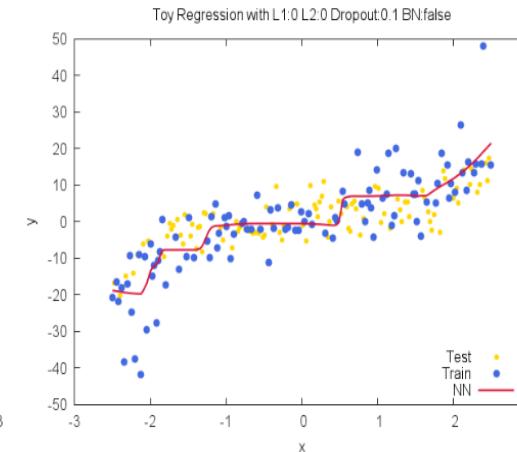
$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{scale and shift}$$

Regularization Methods

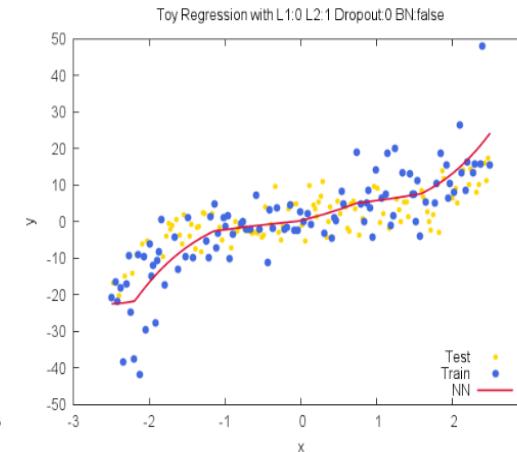
No Regularization



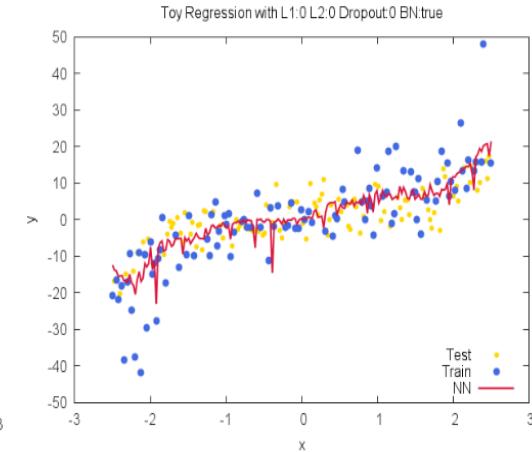
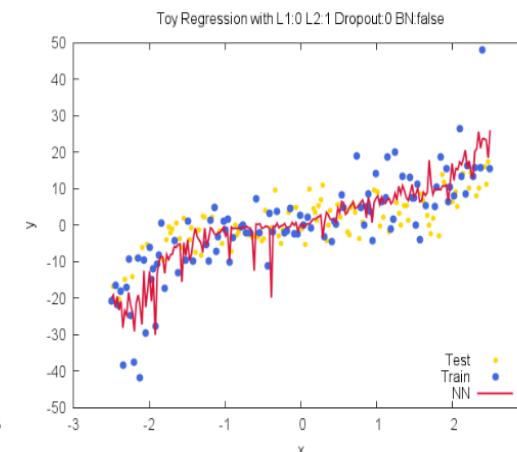
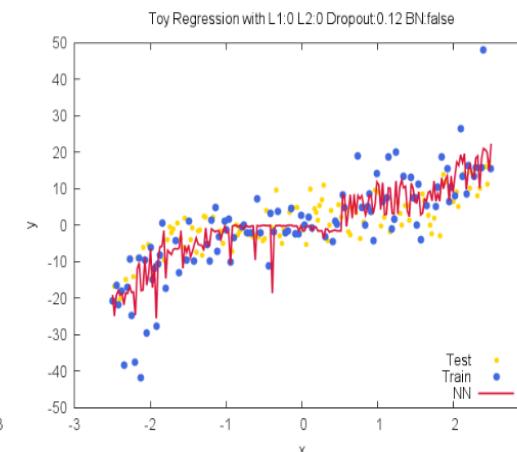
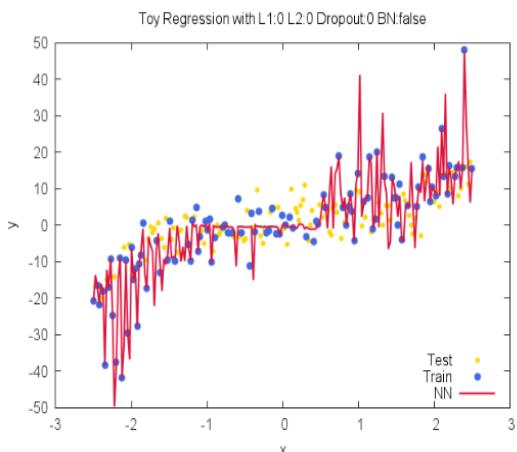
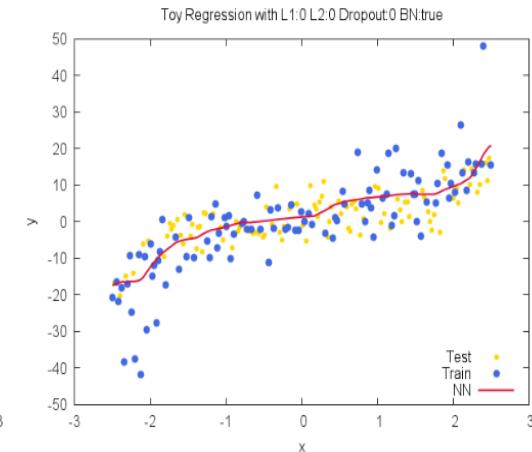
Dropout



L2 Regularization



Batch Normalization

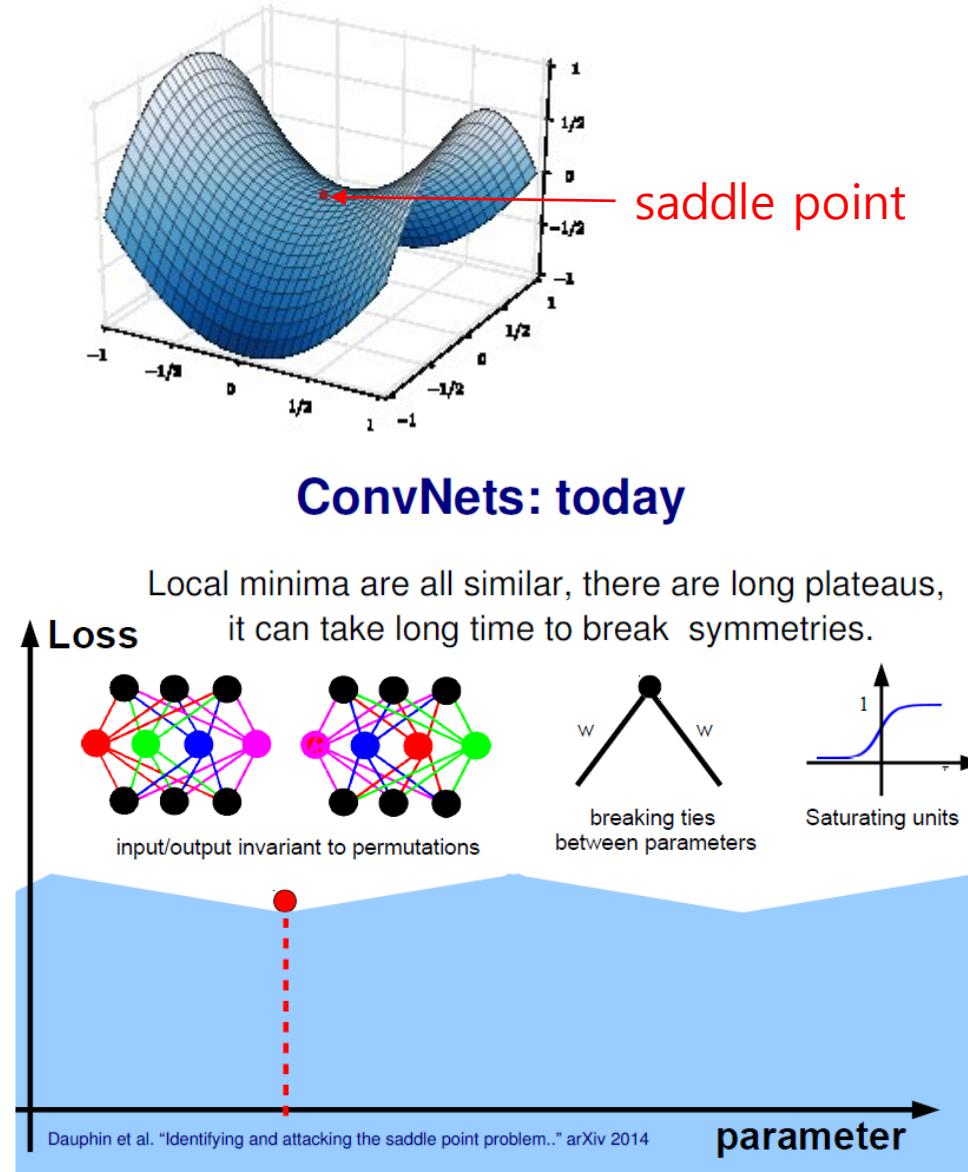


Local Minima

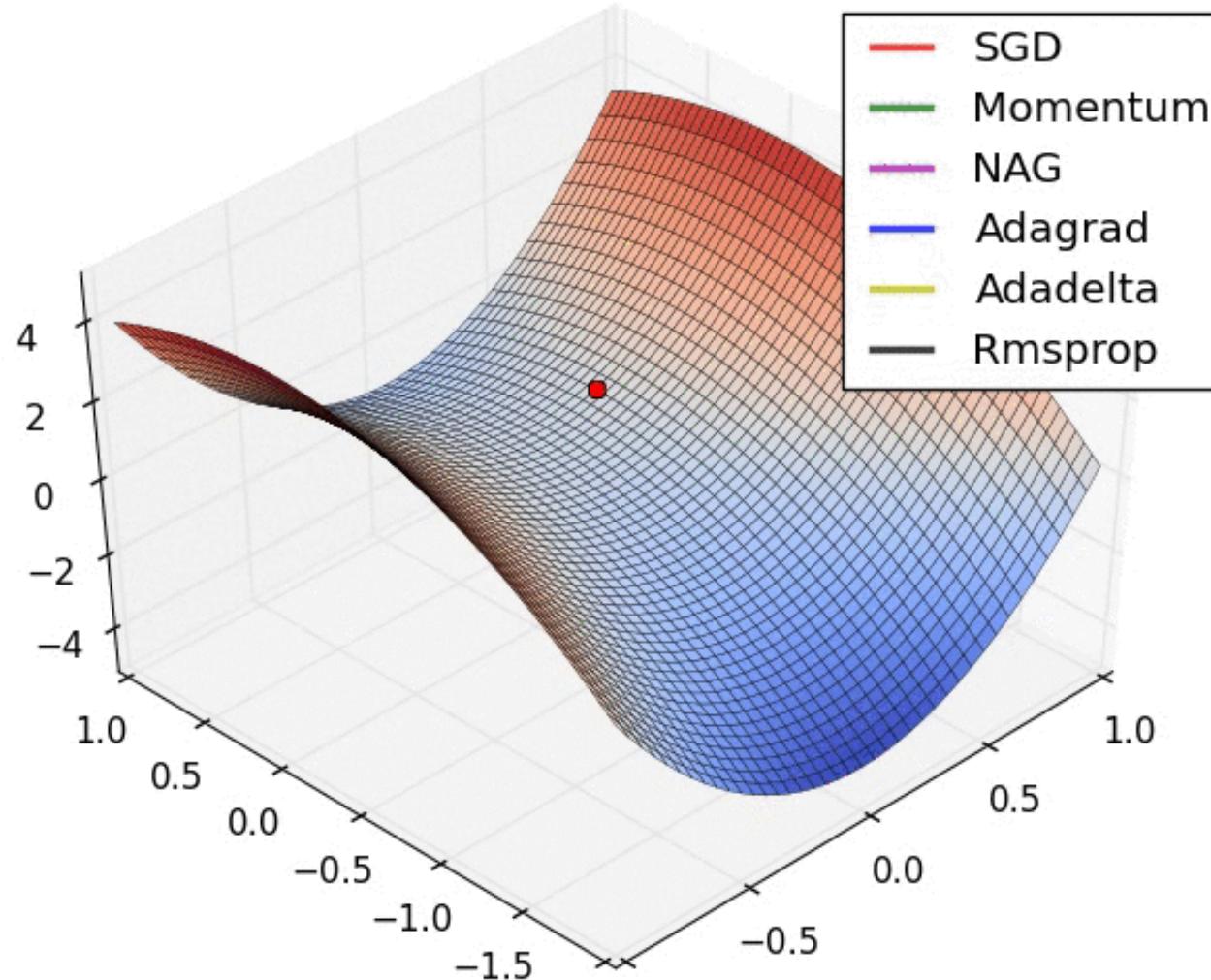


Local Minima에 대하여

- minimum이라고 하는 것은 현재 차원에서 이동할 수 있는 모든 방향으로의 gradient 값이 증가하는 방향이어야 하는데 이런 경우는 확률적으로 희박함
- DNN과 같은 고차원 구조에서는 대부분은 local minima가 아니라 saddle point일 가능성이 높음
- 만약 실제 local minima가 존재한다면 그것은 global minimum과 거의 차이가 없을 가능성이 높음(neural network의 대칭성)



Optimization Methods

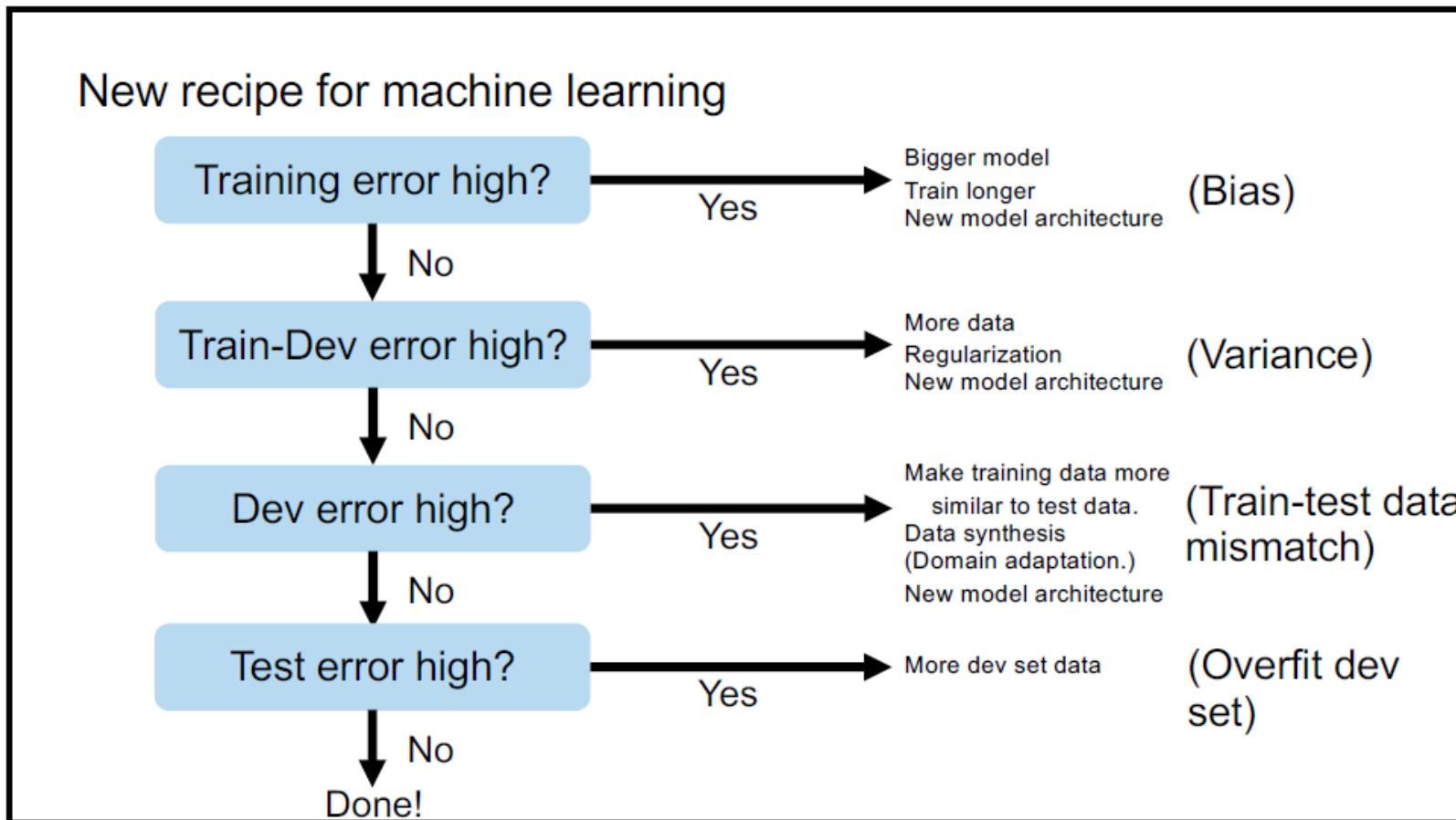


그 밖에 Deep Learning을 가능하게 했던 것들

- Hardware
 - CPUs, GPUs!, ASICs <https://youtu.be/-P28LKWTzrl>
- Organized Large Datasets
 - ImageNet
- Algorithms and Research
 - Backprop, CNN, LSTM
- Software and Infrastructure
 - Git, AWS, Amazon Mechanical Turk, TensorFlow, ...
- Financial Backing of large Companies
 - Google, Facebook, Amazon, ...

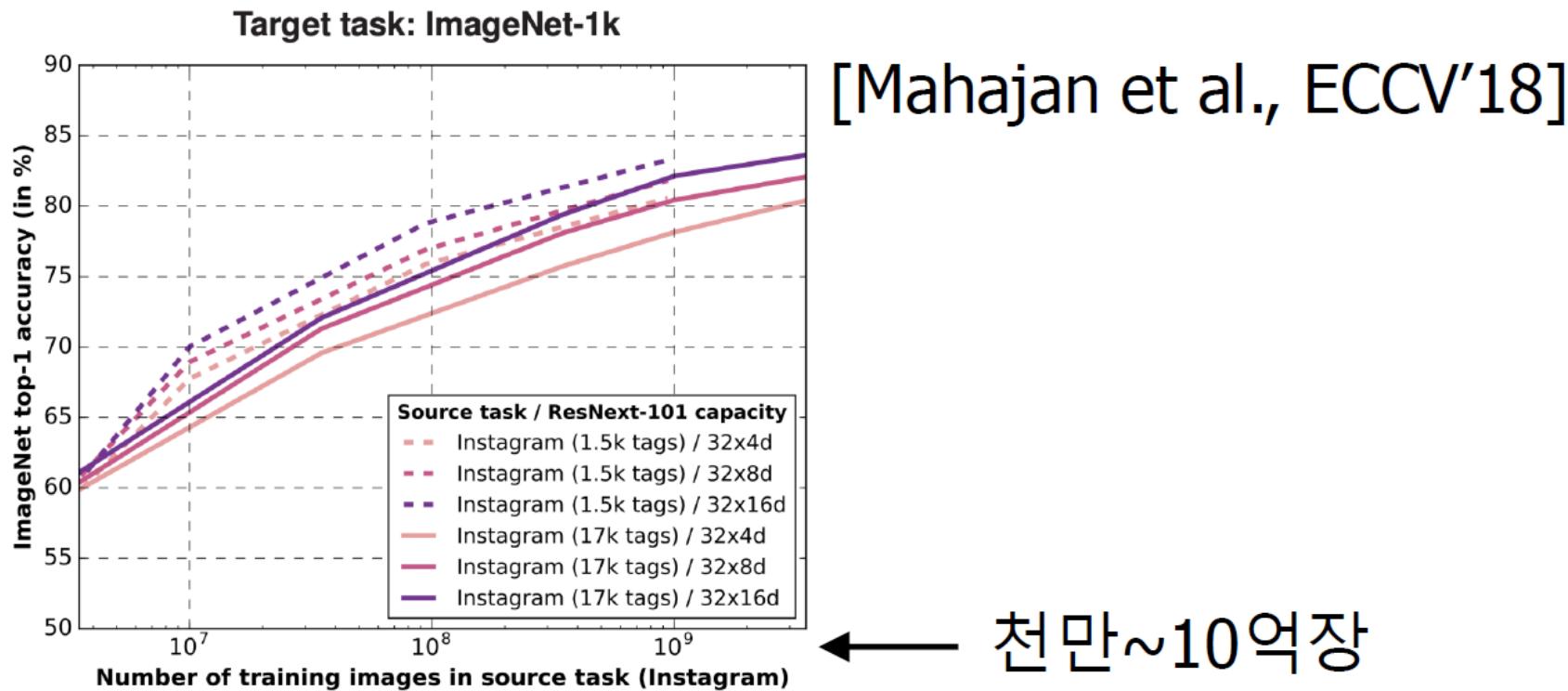
Deep Learning – Very Data Hungry

- Data, more data (by Andrew Ng's tutorial @NIPS 2016)



Data is Important!

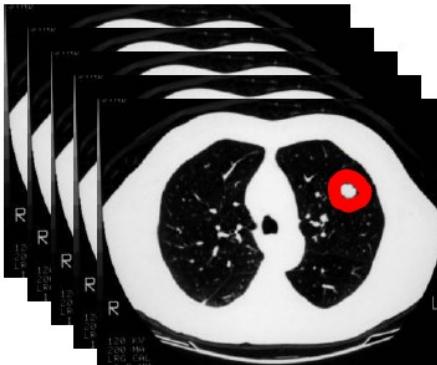
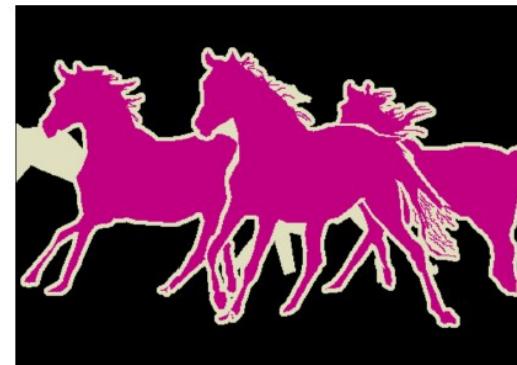
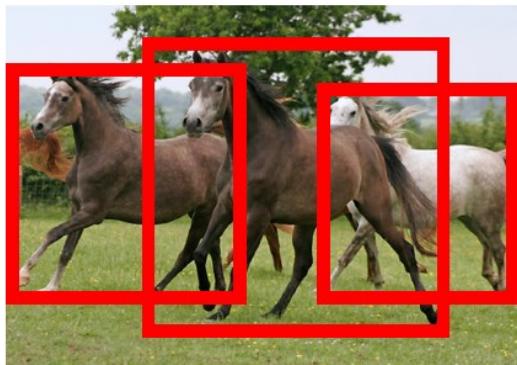
- It is not questionable that
more data still improves network performance



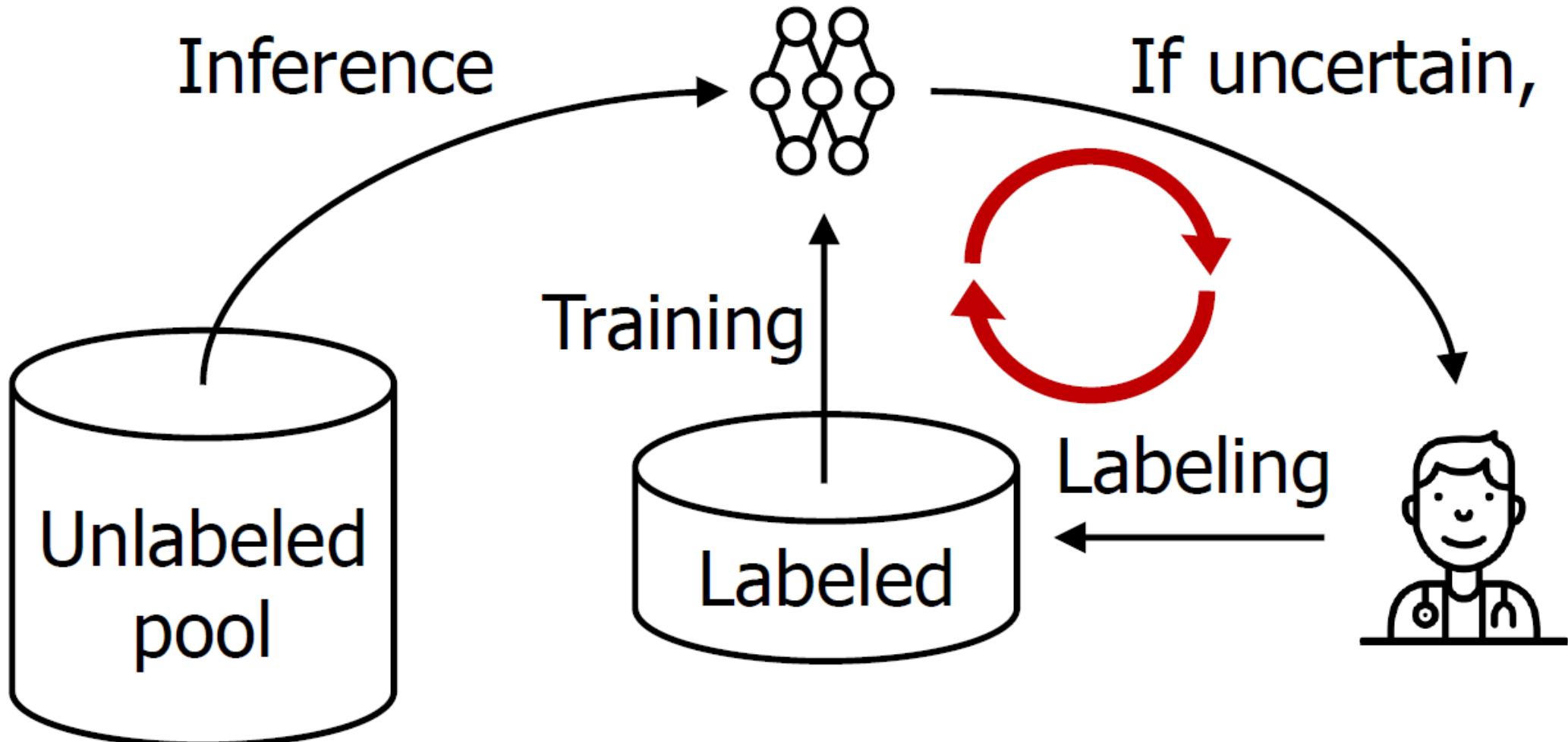
Data is Expensive!



\$

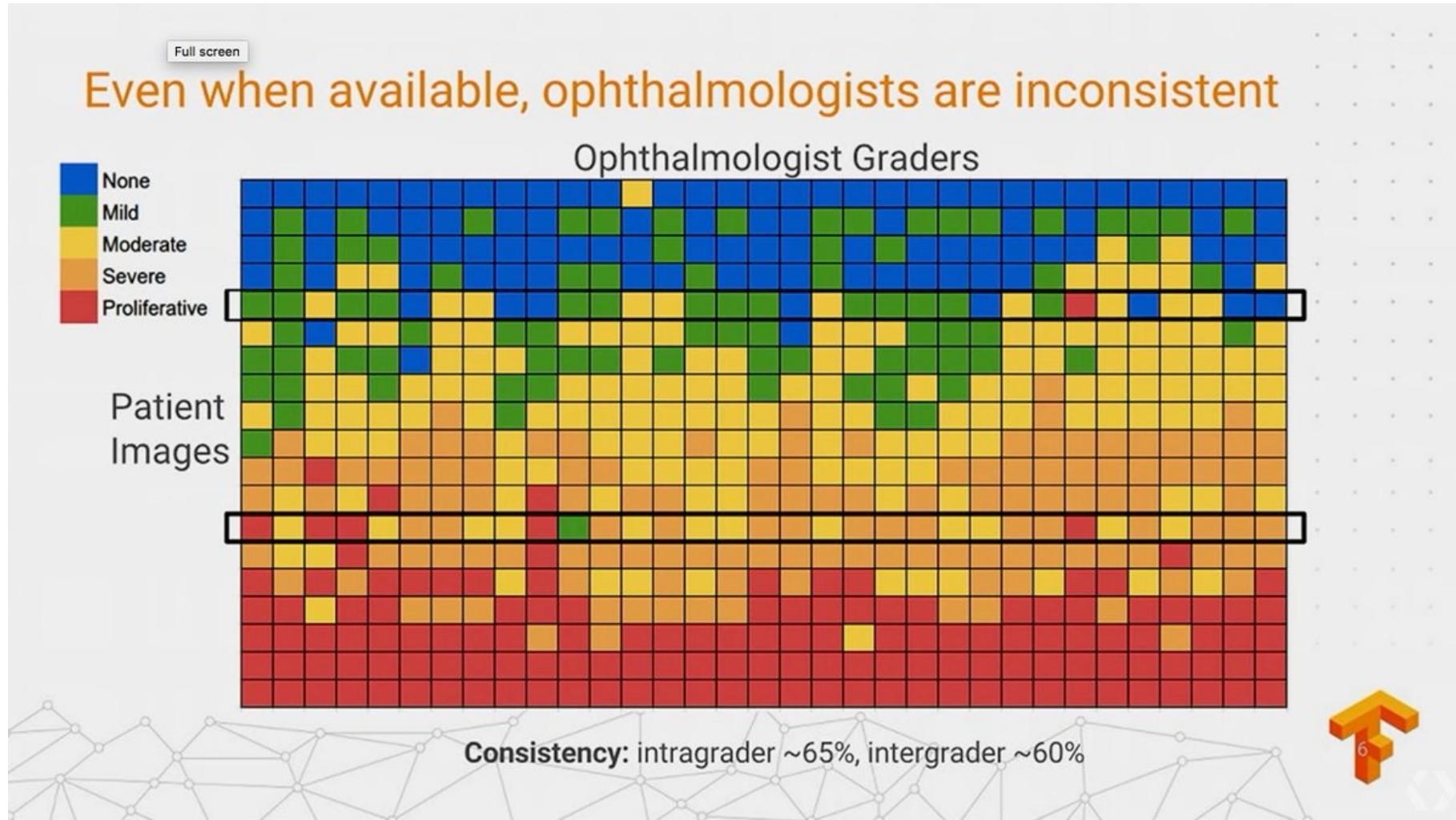


Active Learning



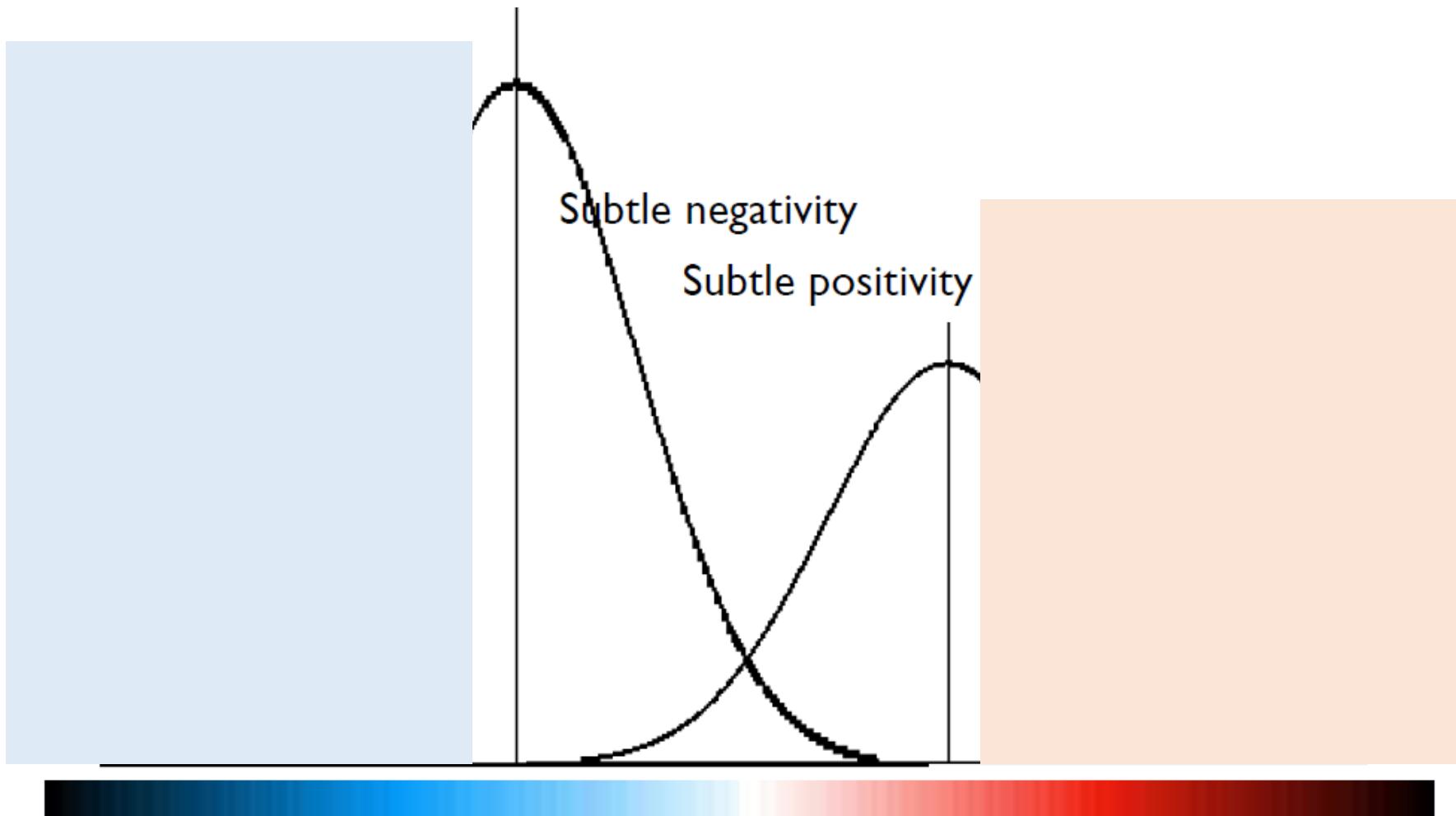
Good Data vs Bad Data

- Our labels are perfect?



Good Data vs Bad Data

- Our data is unbiased?



Is Our Model Good Enough?

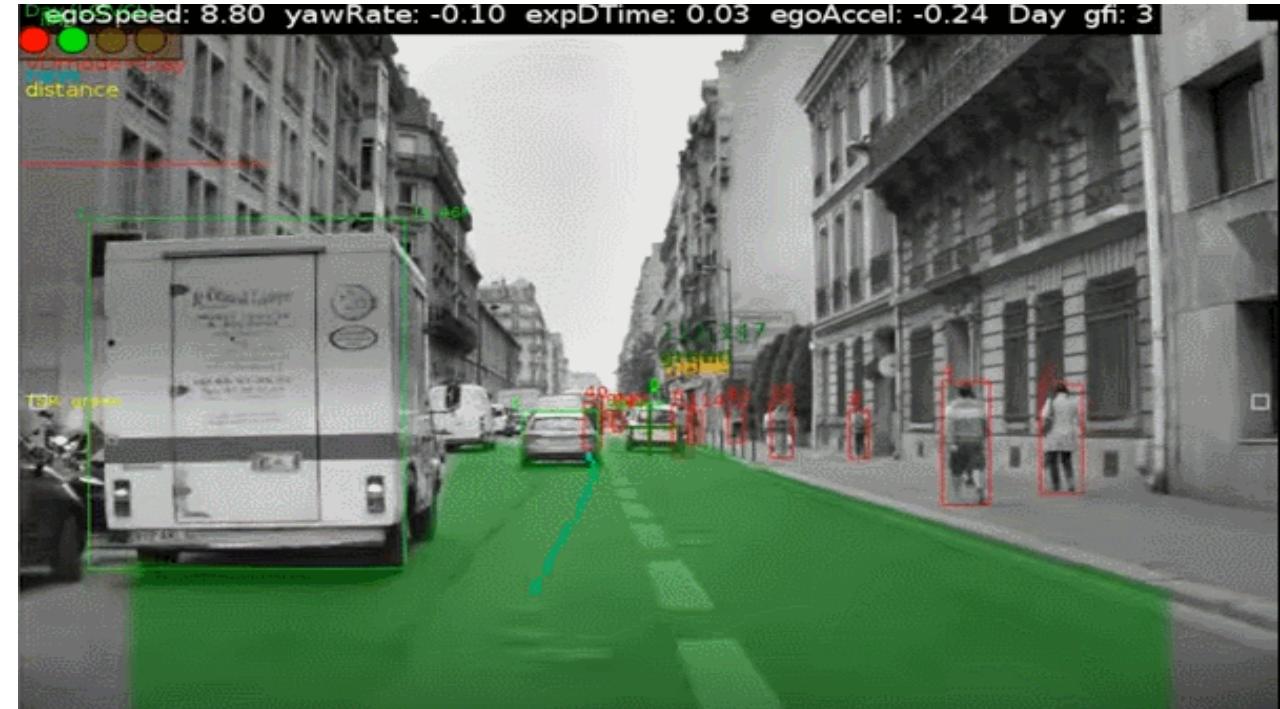
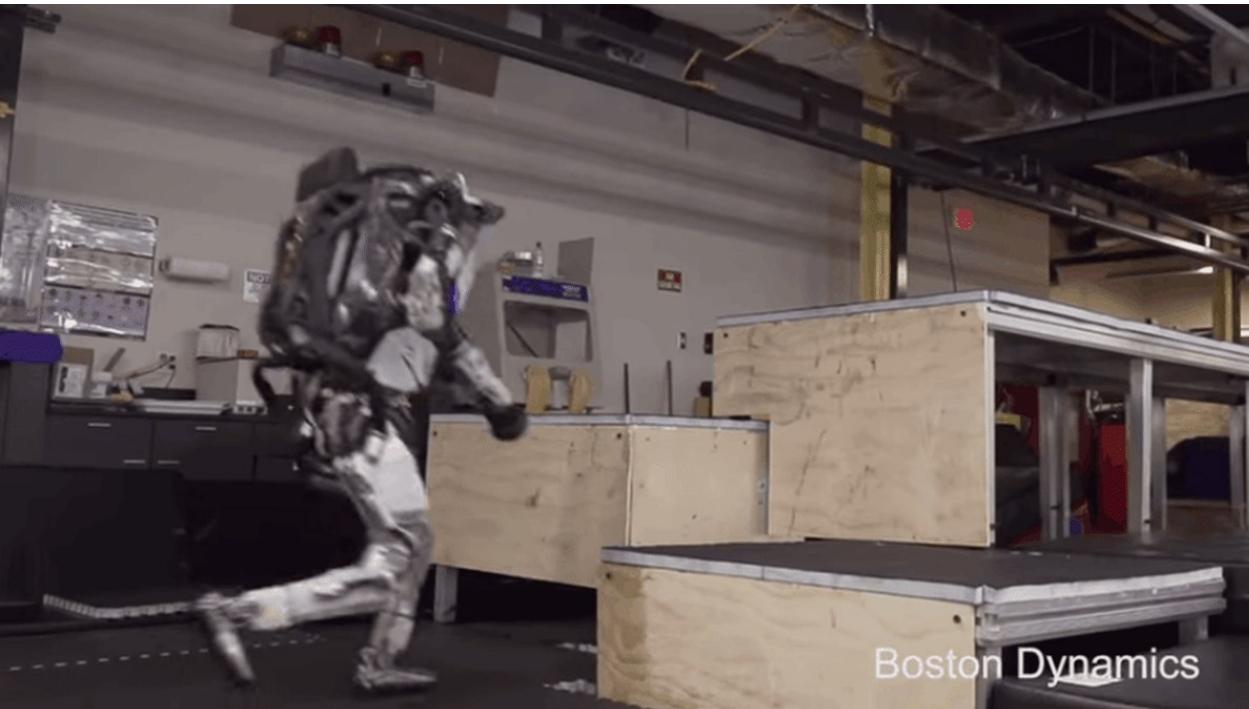
- Sometimes accuracy is not a good measurement

| | | Predicted class | |
|--------------|-----|----------------------|----------------------|
| | | P | N |
| Actual Class | P | True Positives (TP) | False Negatives (FN) |
| | N | False Positives (FP) | True Negatives (TN) |

Sensitivity
Recall

Precision

Deep Learning's Challenges – Toward the Real World

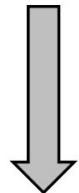


Rethinking Generalization

- Training with random label



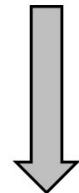
~~dog~~



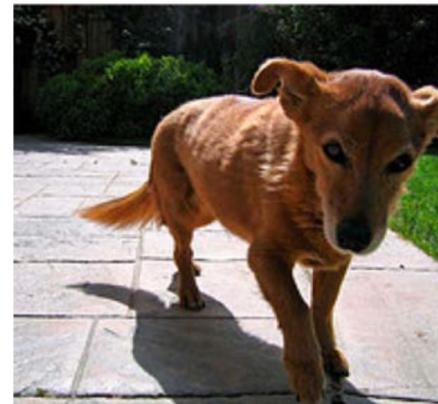
banana



~~banana~~



dog



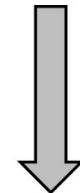
~~dog~~



tree



~~tree~~



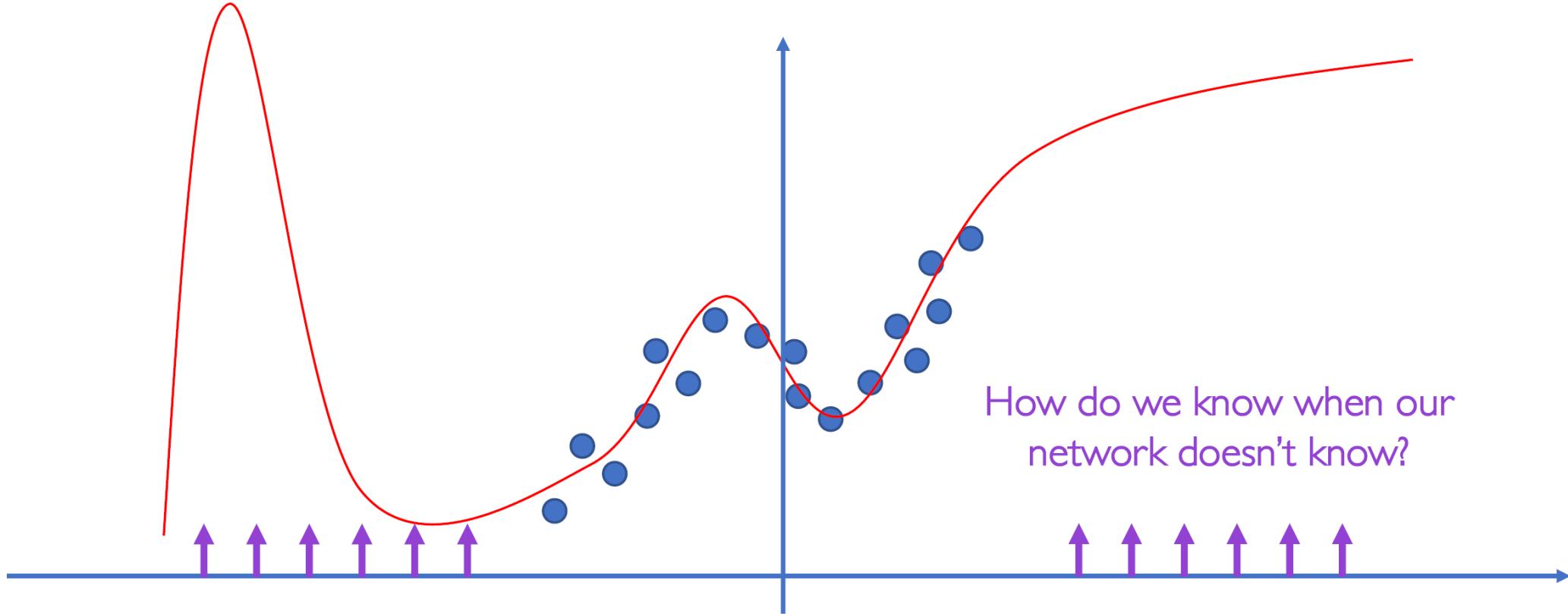
dog

Rethinking Generalization



Interpolation vs Extrapolation

Neural networks are excellent function approximators
...when they have training data



Unintended Consequences

Human Play



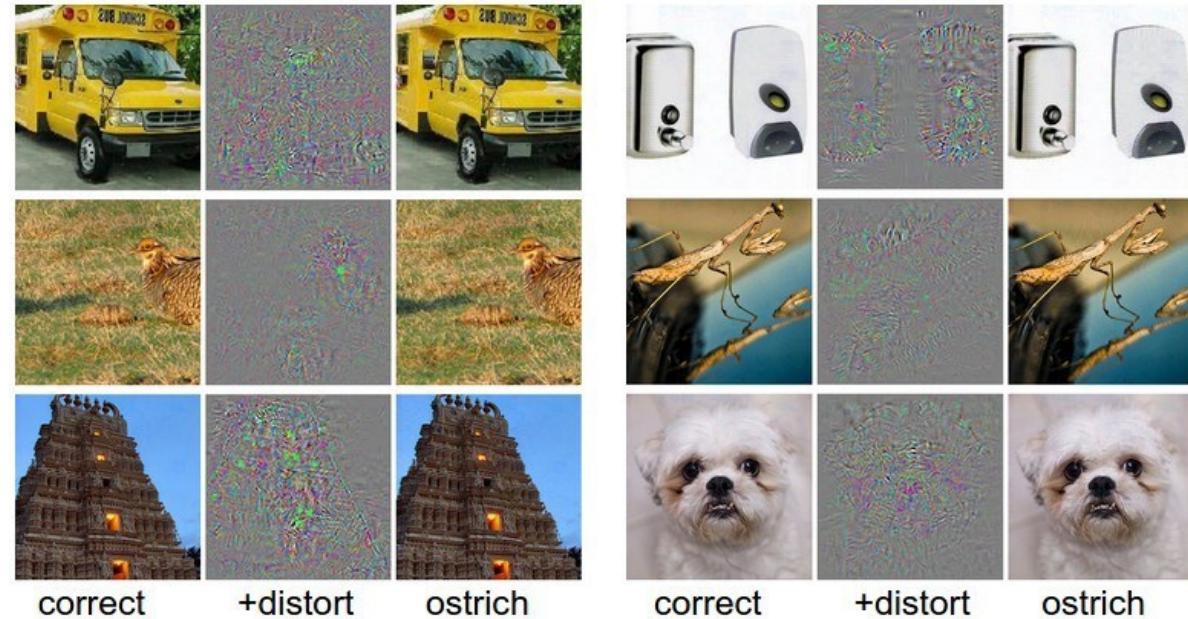
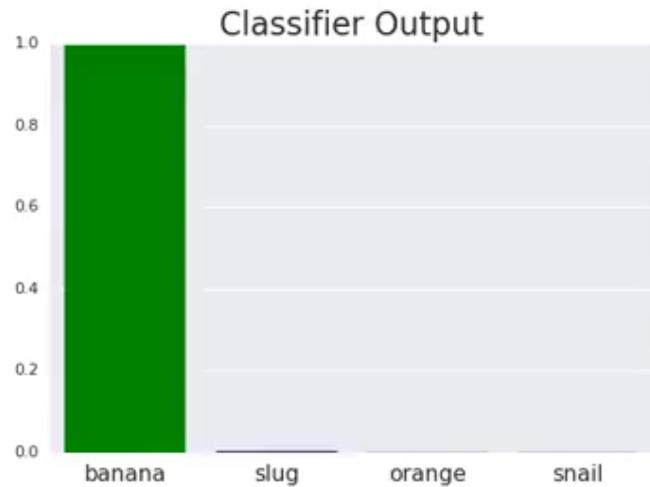
AI Play



Player gets reward based on:

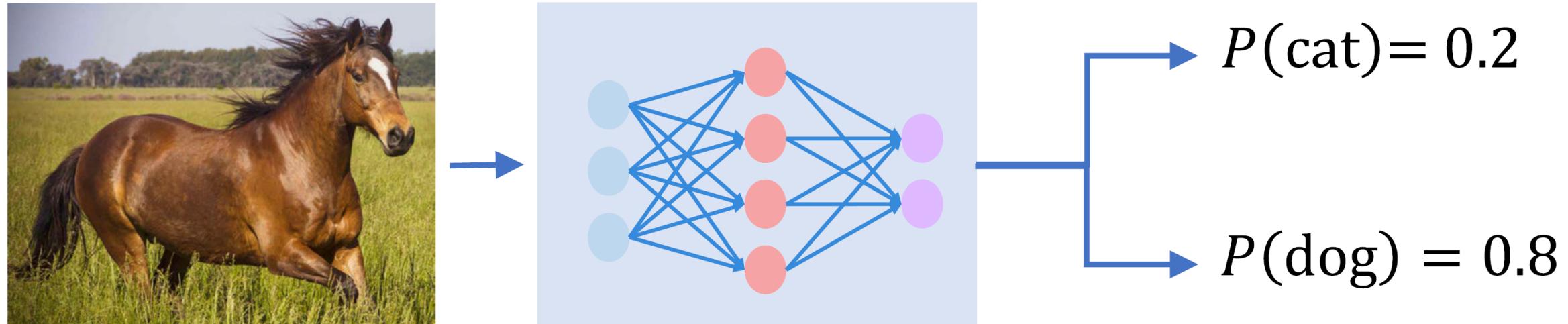
1. Finishing time
2. Finishing position
3. Picking up “turbos”

Adversarial Attack



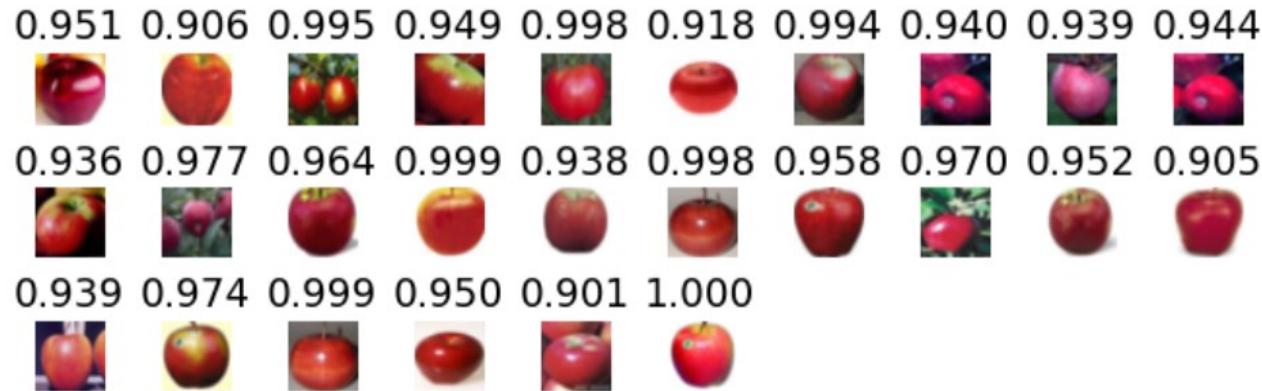
Uncertainty

We need **uncertainty** metrics to assess the network's **confidence** in its predictions.

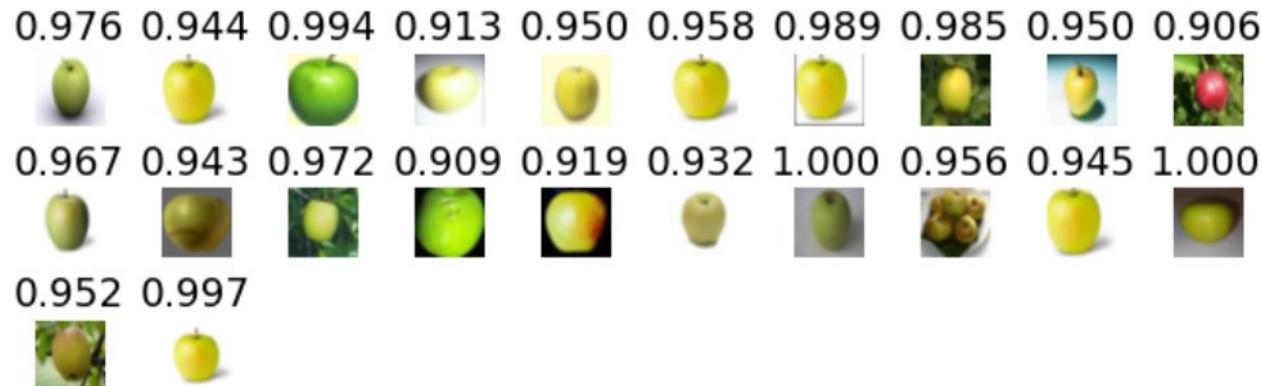


Remember: $P(\text{cat}) + P(\text{dog}) = 1$

Uncertainty

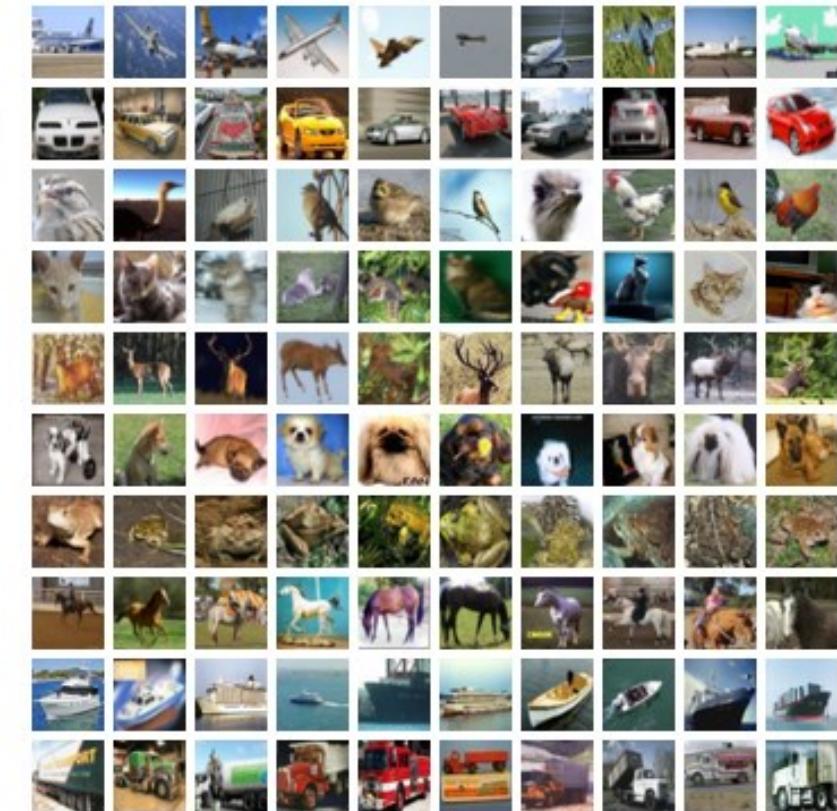


CIFAR-100's *apple* misclassified as CIFAR-10's *automobile* class with
 $p > 0.9$.



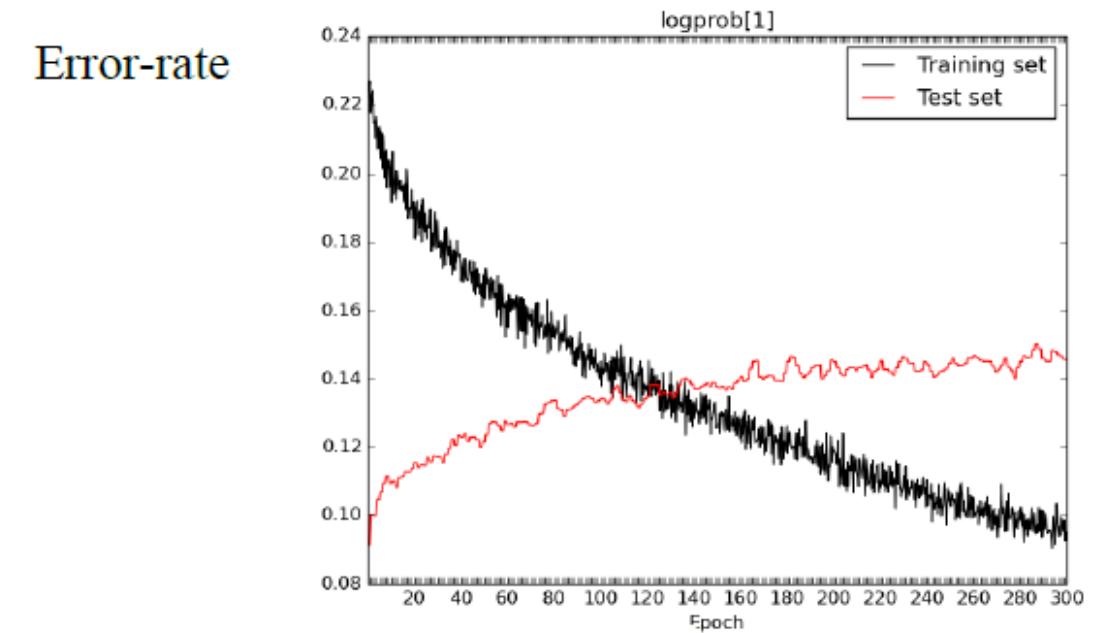
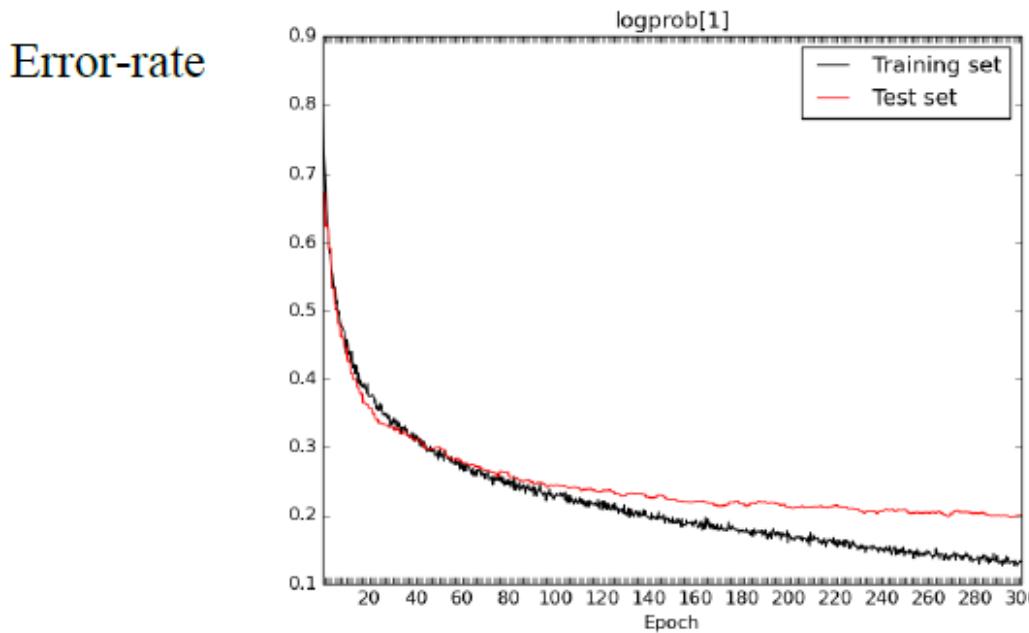
CIFAR-100's *apple* misclassified as CIFAR-10's *frog* class with $p > 0.9$.

Cifar 10 dataset

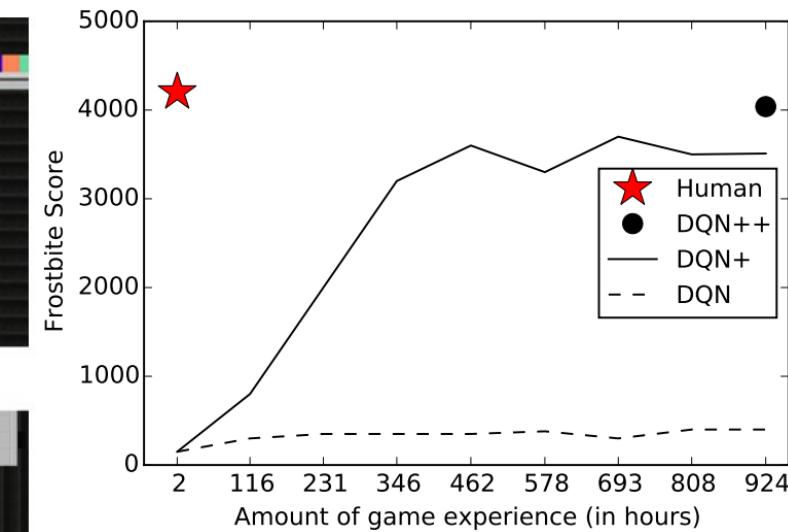
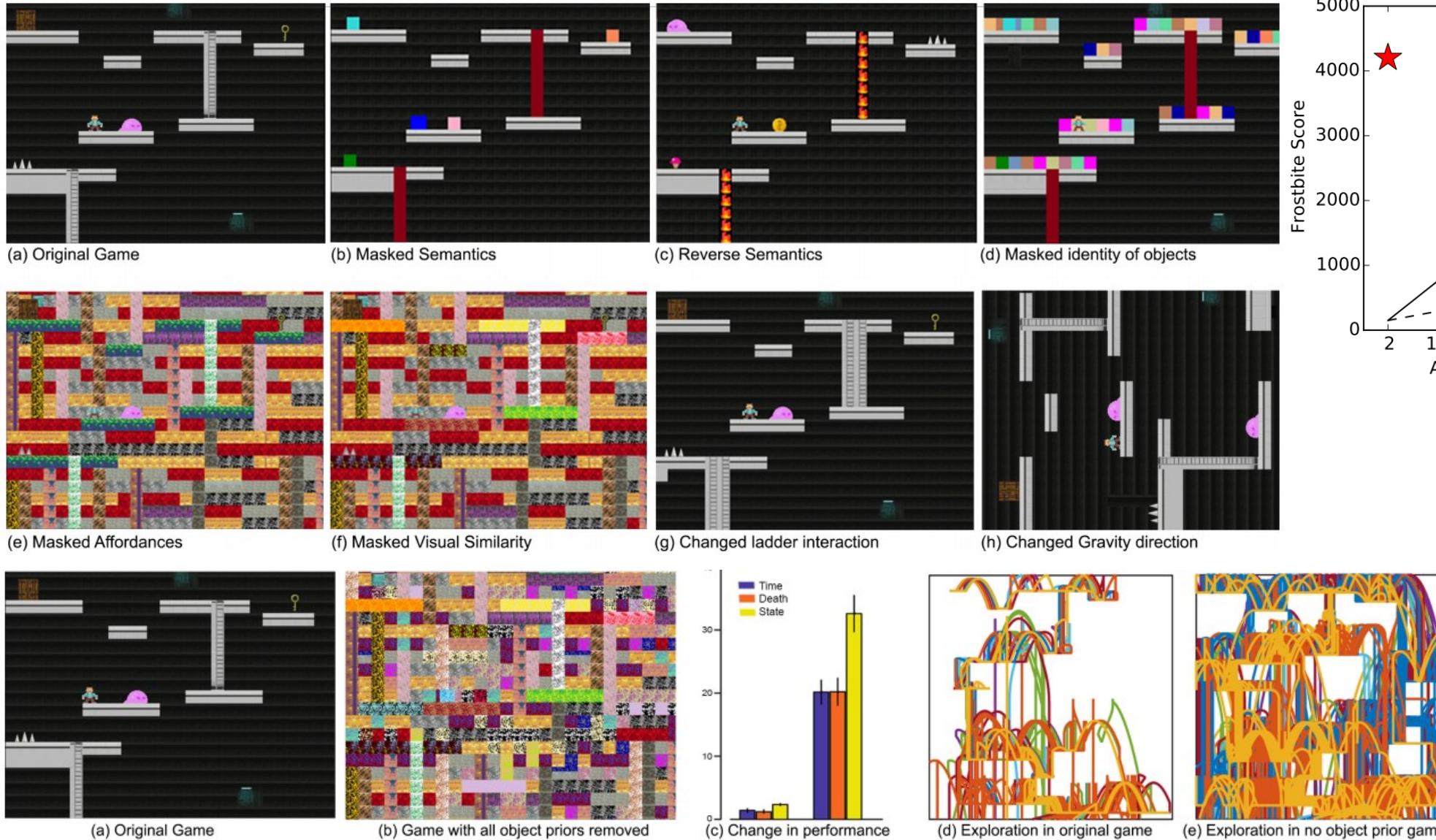


Catastrophic Forgetting Problem

- 학습 데이터를 30,000(group 1), 30,000(group 2)로 구성
- Group 1로 network 학습(test는 group 2)
- 학습된 weights를 initial weights로 하여 group 2를 다시 학습



Inefficient Training



Limitations of Deep Learning

- Not sample efficient
- Compute intensive
- Poor at handling prior knowledge or uncertainty
- Not interpretable
- Suffers from hyper-parameter optimization
- Vulnerable to adversarial attack

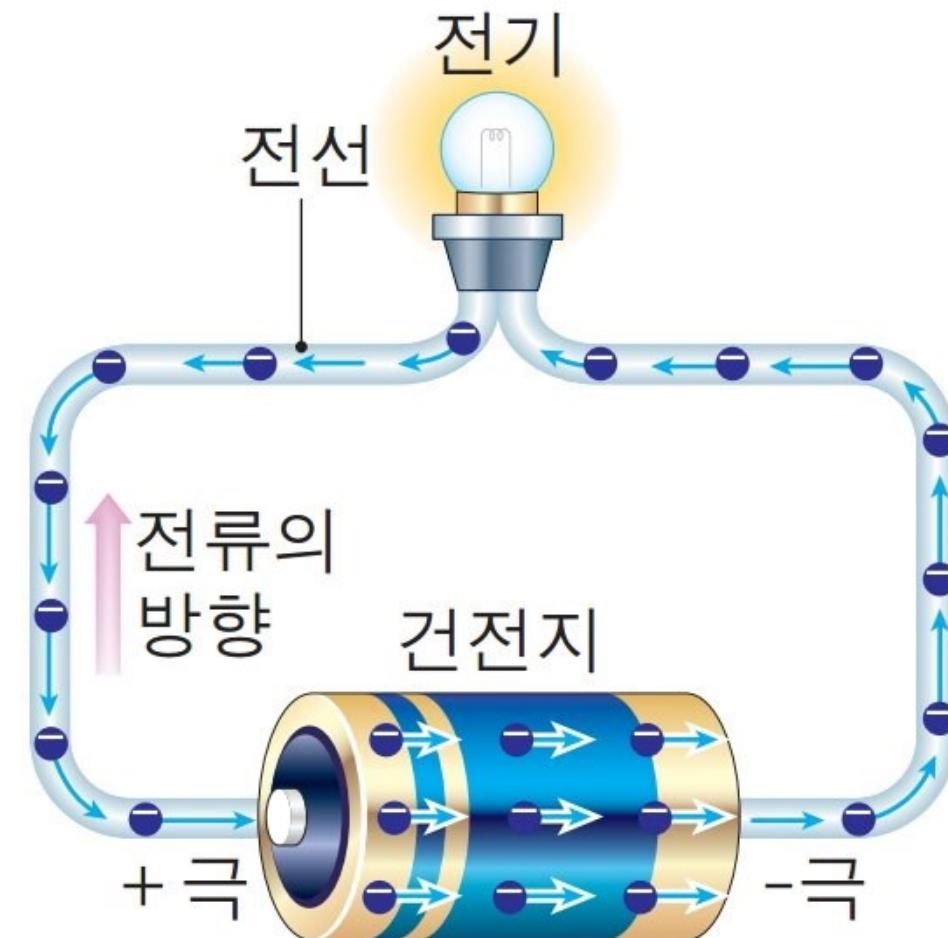
Hot Topics of Deep Learning(2020)

- Graph Neural Network
- Meta Learning
- Attention (in NLP)
- Self-supervised Learning
- Adversarial Attack & Training
- Auto-ML
- Model Compression
- ...

Today's Artificial Intelligence?



L'Avion III de Clément Ader, 1897
(Musée du CNAM, Paris)



▲ 전자의 이동과 전류의 방향