딥러닝을 이용한 보안

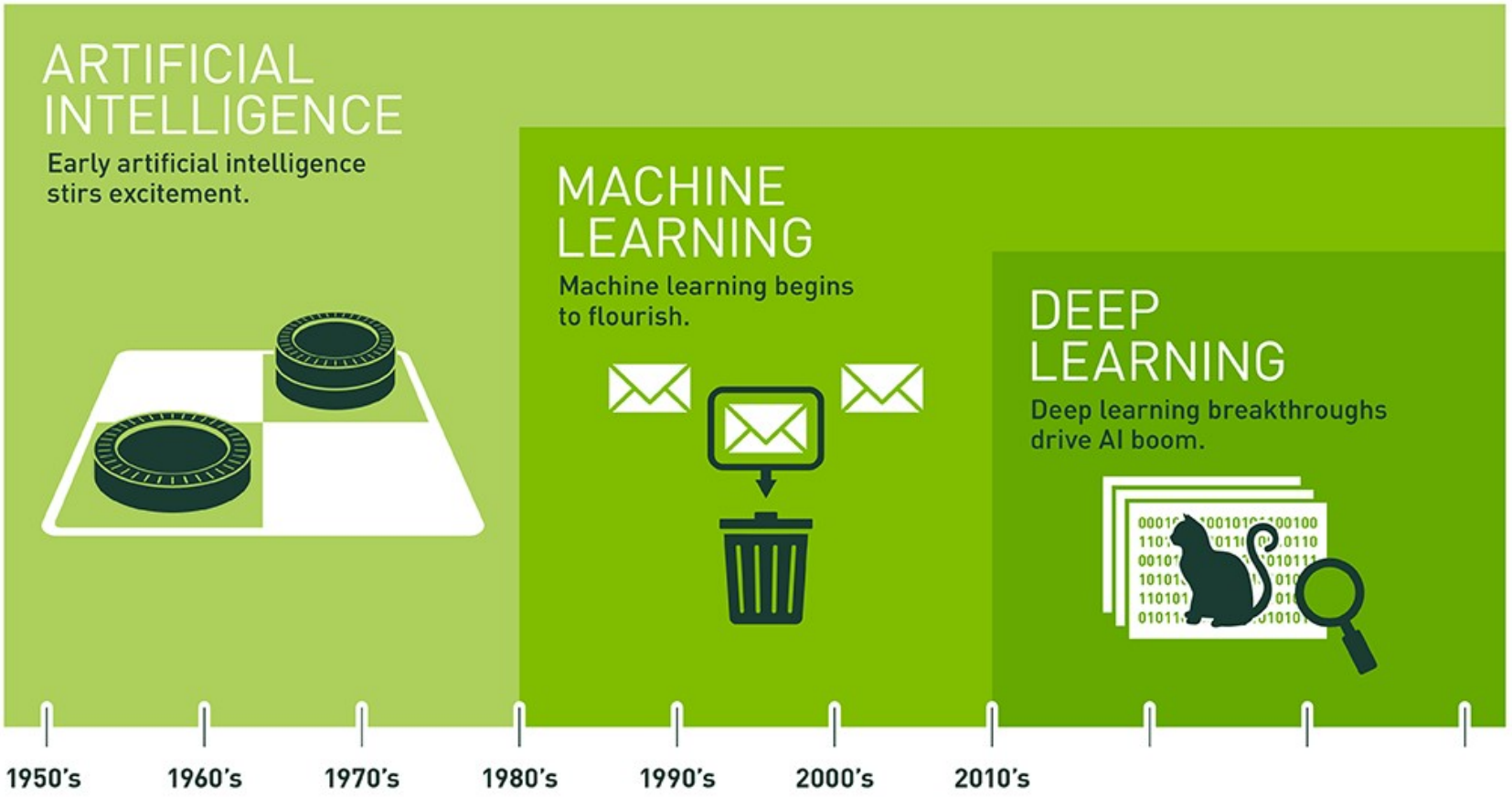
최승우

한국과학기술원 지식서비스공학대학원, 박사과정

Contents

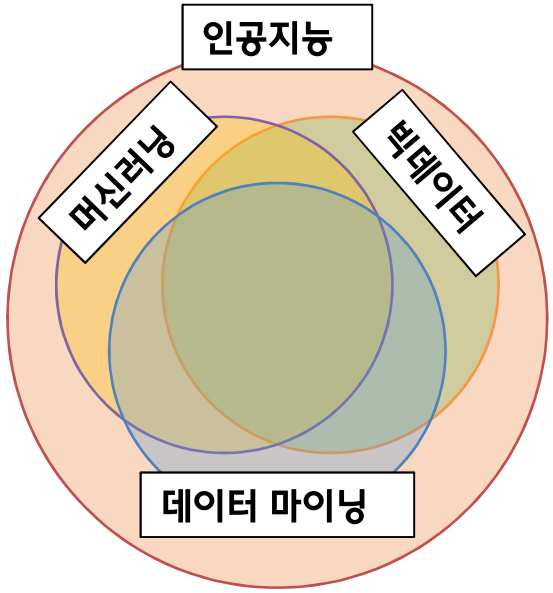
Contents

인공지능, 기계학습, 딥러닝?!

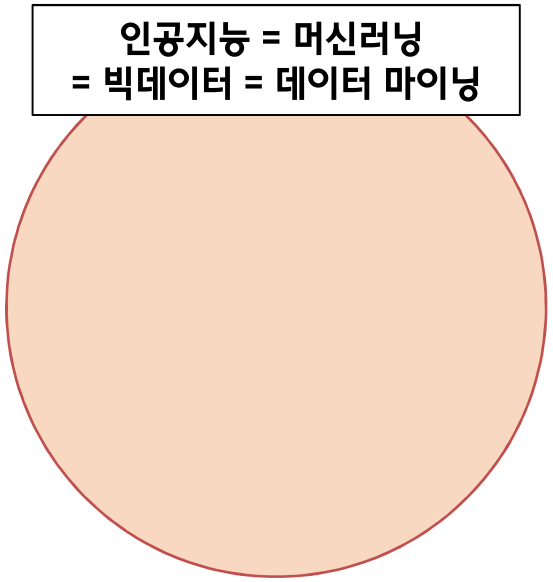


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

인공지능, 기계학습, 딥러닝?!

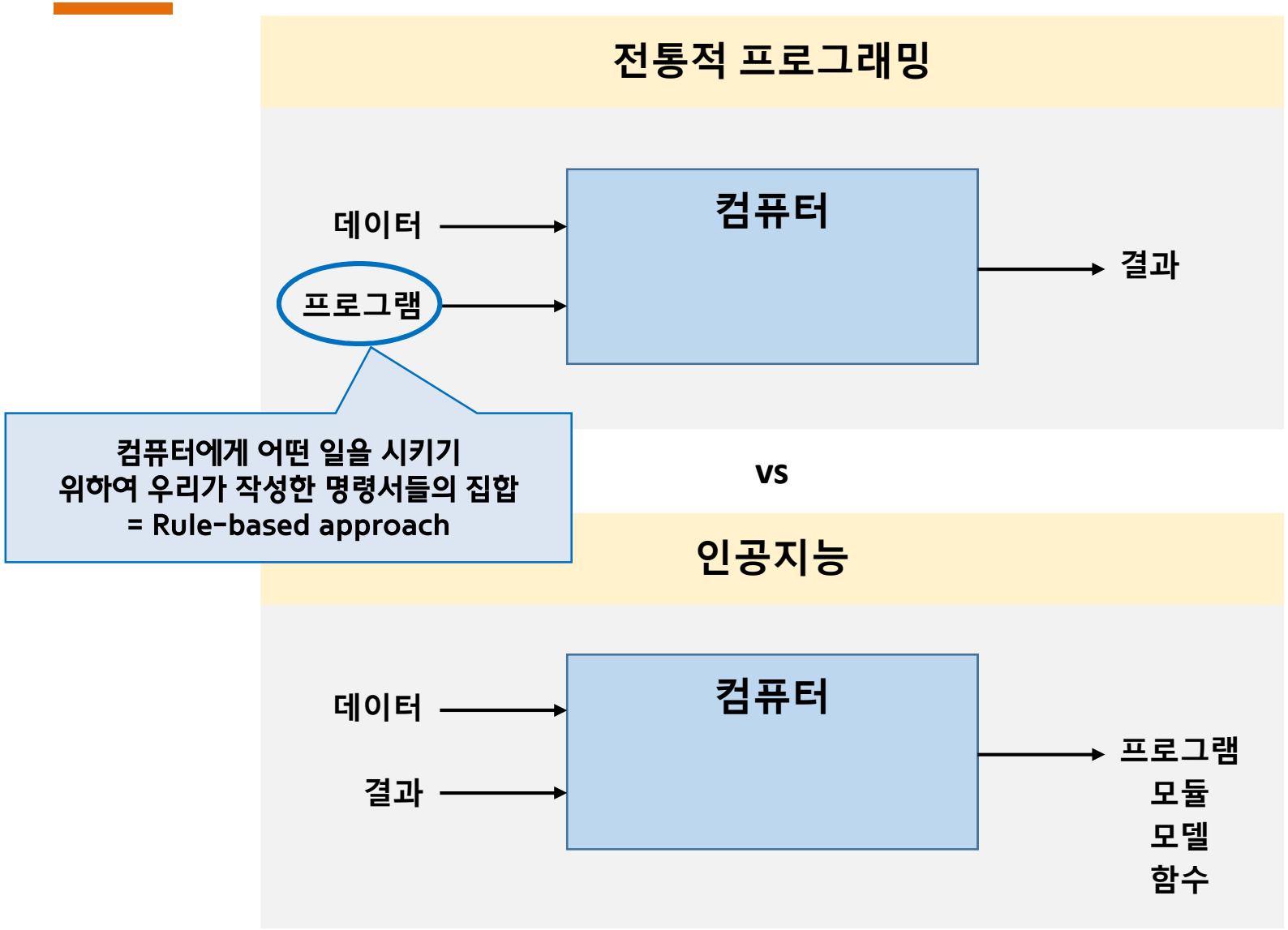


현실



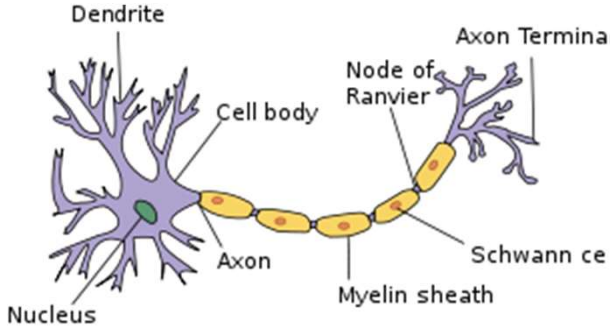
이상

인공지능, 기계학습, 딥러닝?!

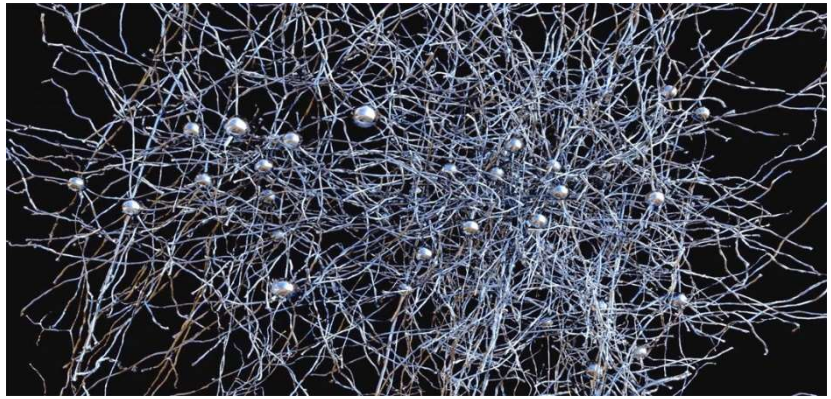


Contents

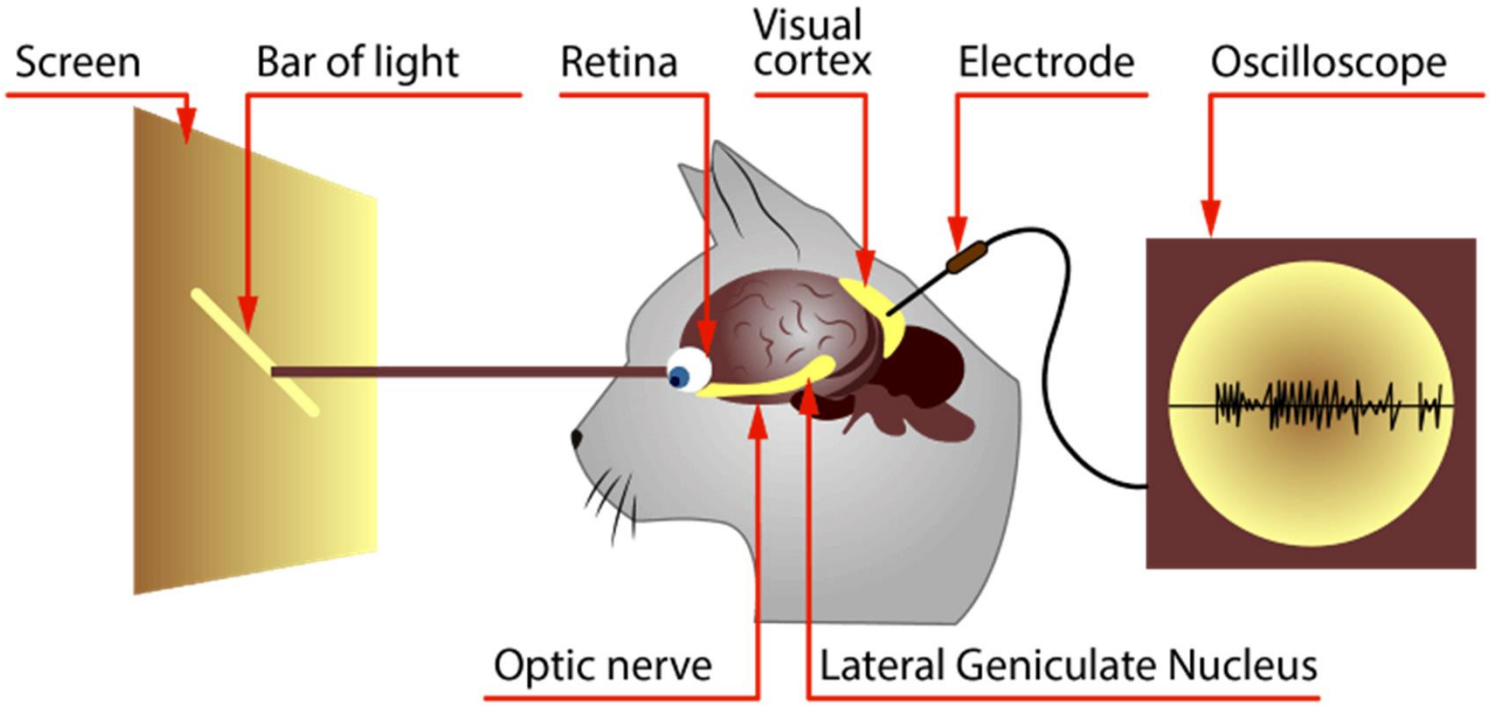
딥러닝 소개 – 정보를 처리하는 인간의 뇌



뉴런



딥러닝 소개 – 뇌에서 정보를 추상화 하는 방법



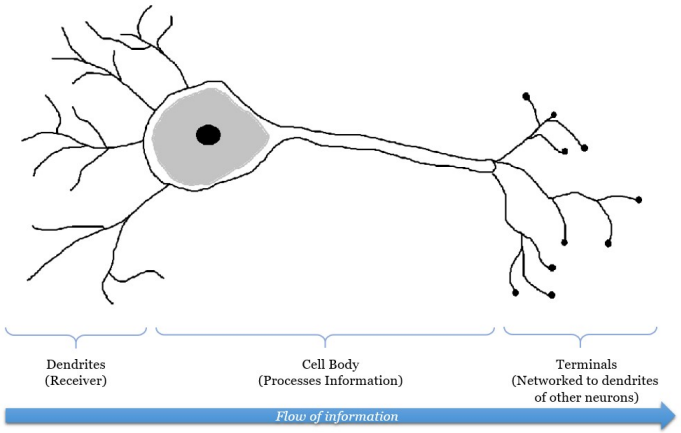
딥러닝 소개 - 요약

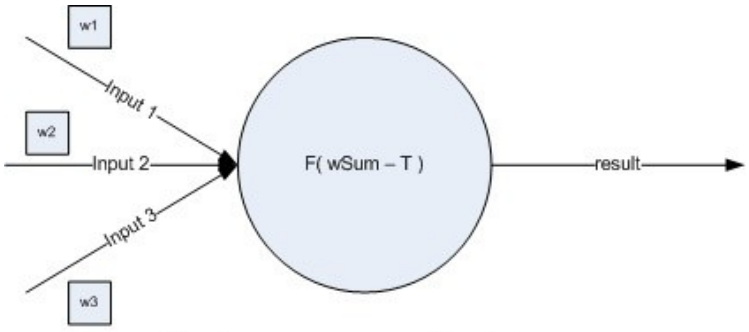
- 뇌의 뉴런을 모사한 인공 뉴런을 구성하여 정보를 추상화 하는 기법.
- 복잡한 정보들의 깊은 (Deep) 정보를 잘 추상화 한다고 알려짐.
- 컴퓨팅 능력의 향상 및 분석할 데이터 양 (Volume)이 증가하면서 최근에 급격하게 발전 하였음.
- 자율주행 (이미지 처리), 챗봇(언어 처리) 등 비정형 데이터들을 처리하는데 많이 활용이 됨.

※ 딥러닝 기법은 인공지능 알고리즘들 중 대표적인 하나의 알고리즘.
다른 알고리즘 : Decision tree, Support Vector Machine (SVM) 등

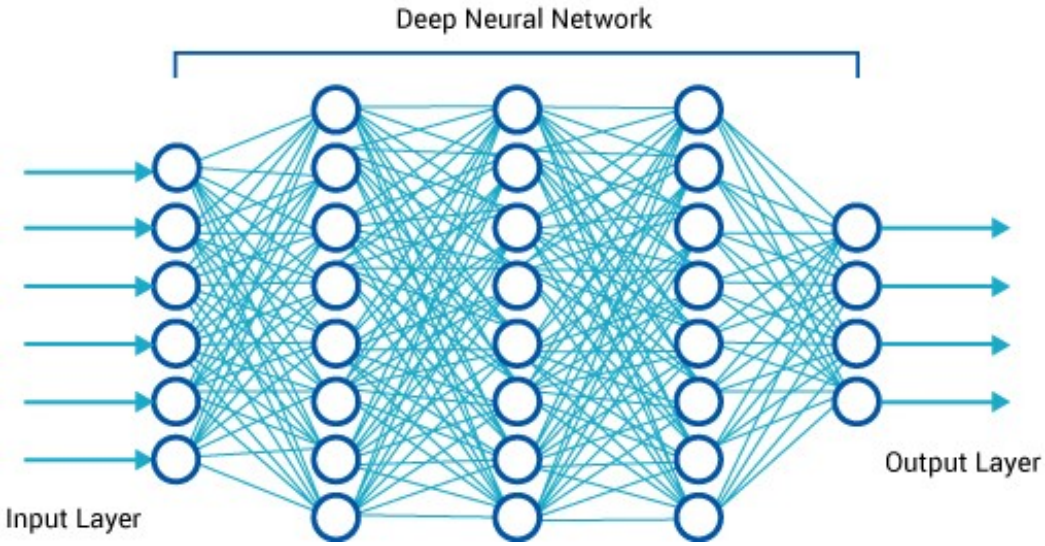
Contents

딥러닝 원리





Perceptron

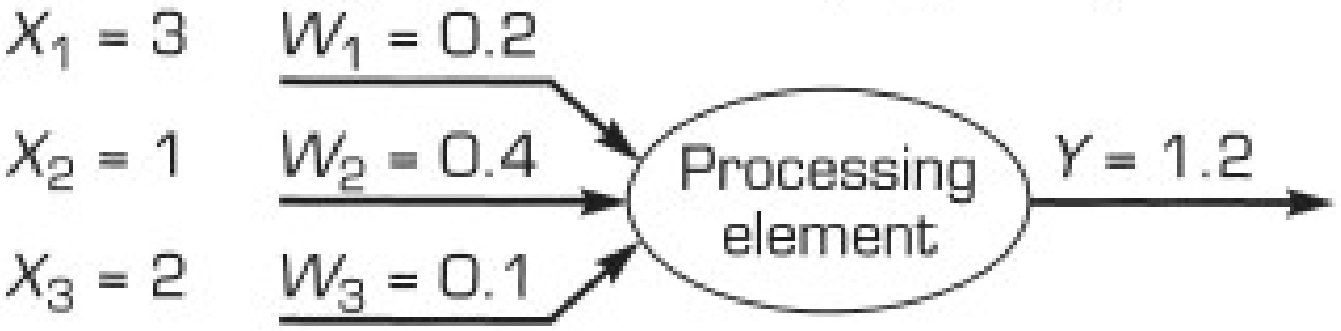


Multi layer perceptron(MLP)

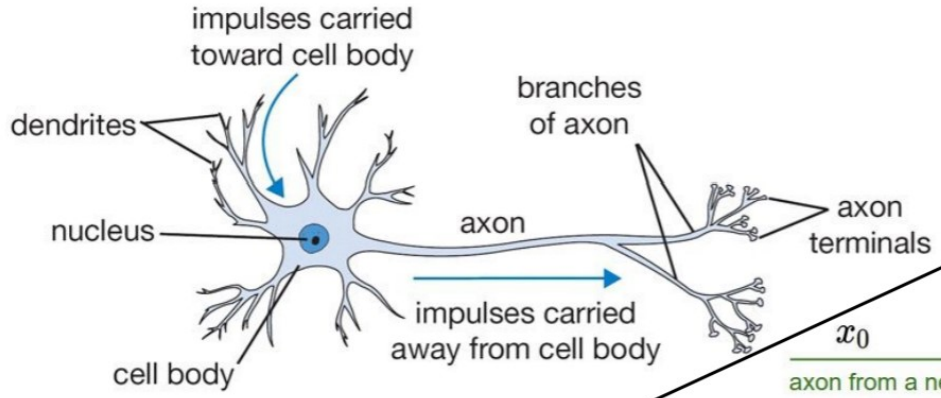
딥러닝 원리

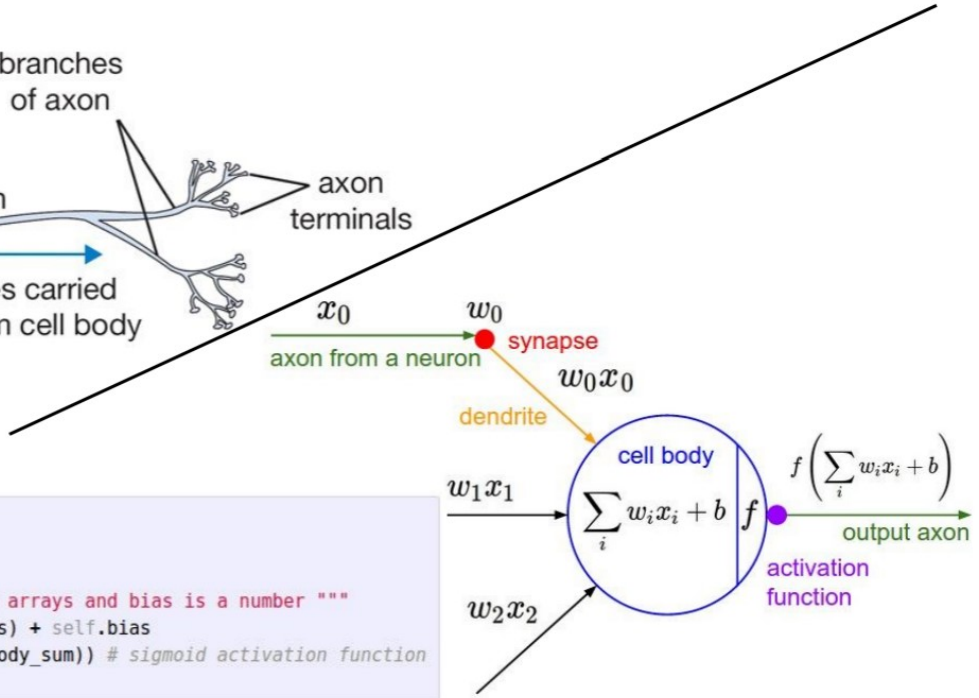
Summation function: $Y = 3(0.2) + 1(0.4) + 2(0.1) = 1.2$

Transformation (transfer) function: $Y_T = 1/(1 + e^{-1.2}) = 0.77$



딥러닝 원리






```
class Neuron:
  # ...
  def neuron_tick(inputs):
    """ assume inputs and weights are 1-D numpy arrays and bias is a number """
    cell_body_sum = np.sum(inputs * self.weights) + self.bias
    firing_rate = 1.0 / (1.0 + math.exp(-cell_body_sum)) # sigmoid activation function
    return firing_rate
```

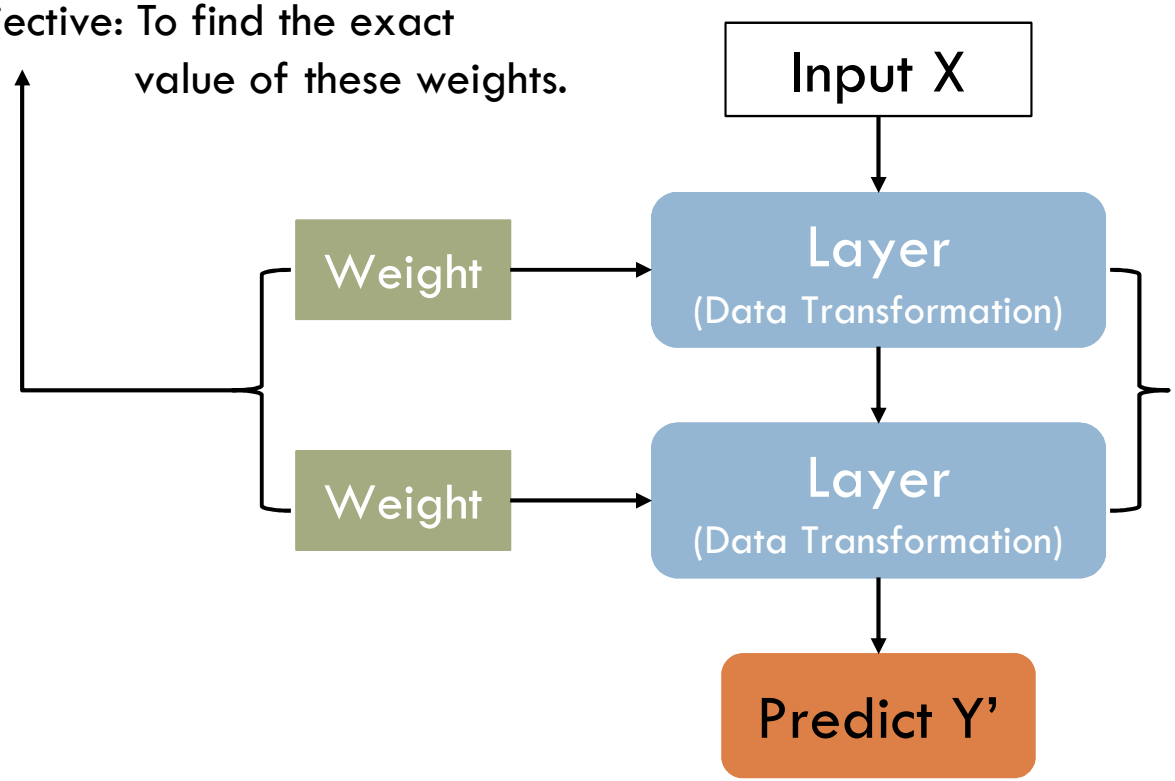
$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix} = \begin{pmatrix} x_{11}w_{11} + x_{12}w_{21} + x_{13}w_{31} & x_{11}w_{12} + x_{12}w_{22} + x_{13}w_{32} \\ x_{21}w_{11} + x_{22}w_{21} + x_{23}w_{31} & x_{21}w_{12} + x_{22}w_{22} + x_{23}w_{32} \\ x_{31}w_{11} + x_{32}w_{21} + x_{33}w_{31} & x_{31}w_{12} + x_{32}w_{22} + x_{33}w_{32} \\ x_{41}w_{11} + x_{42}w_{21} + x_{43}w_{31} & x_{41}w_{12} + x_{42}w_{22} + x_{43}w_{32} \\ x_{51}w_{11} + x_{52}w_{21} + x_{53}w_{31} & x_{51}w_{12} + x_{52}w_{22} + x_{53}w_{32} \end{pmatrix}$$

[n, 3] **[3, 2]** [n, 2]

$$H(X) = XW$$

딥러닝 원리

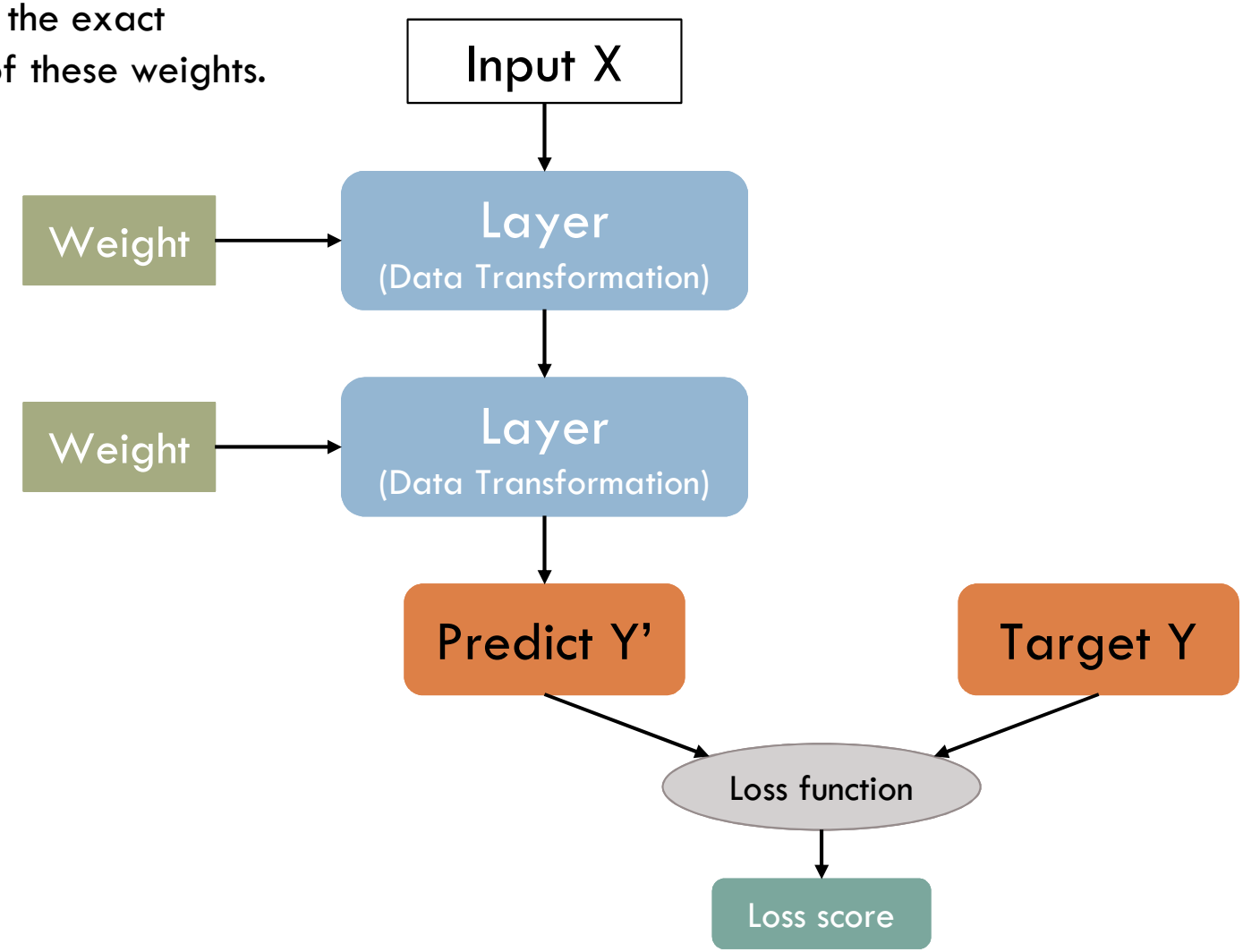
Objective: To find the exact value of these weights.



These layers can be considered as artificial neurons in previous slide.

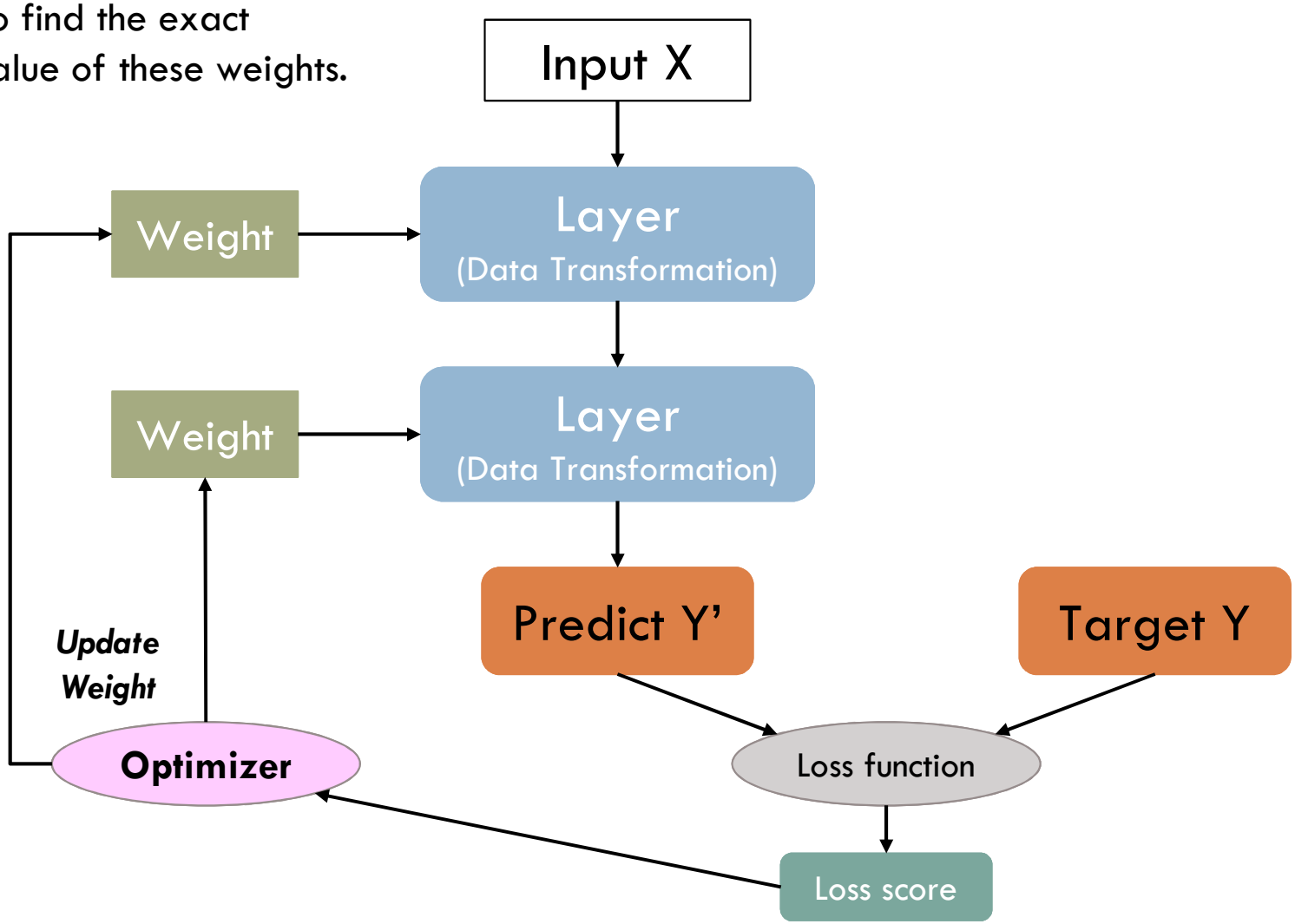
딥러닝 원리

Objective: To find the exact value of these weights.

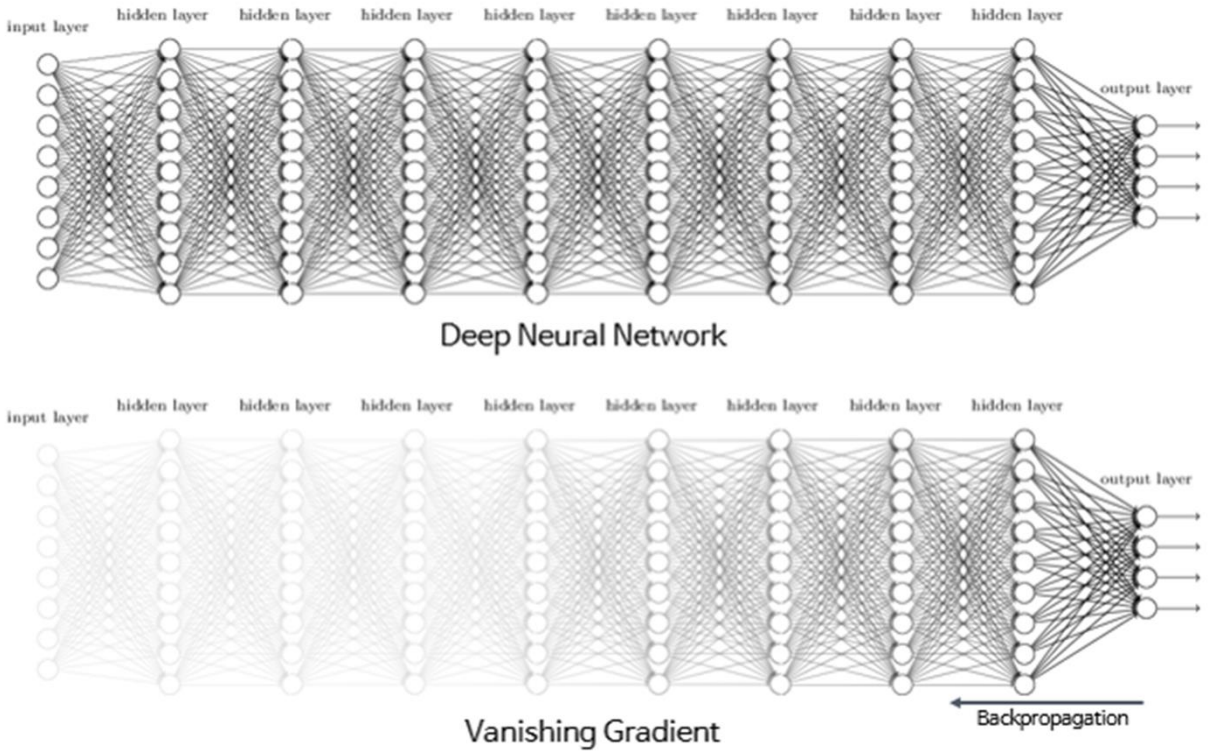


딥러닝 원리

Objective: To find the exact value of these weights.



딥러닝 원리 – Vanishing gradient 문제



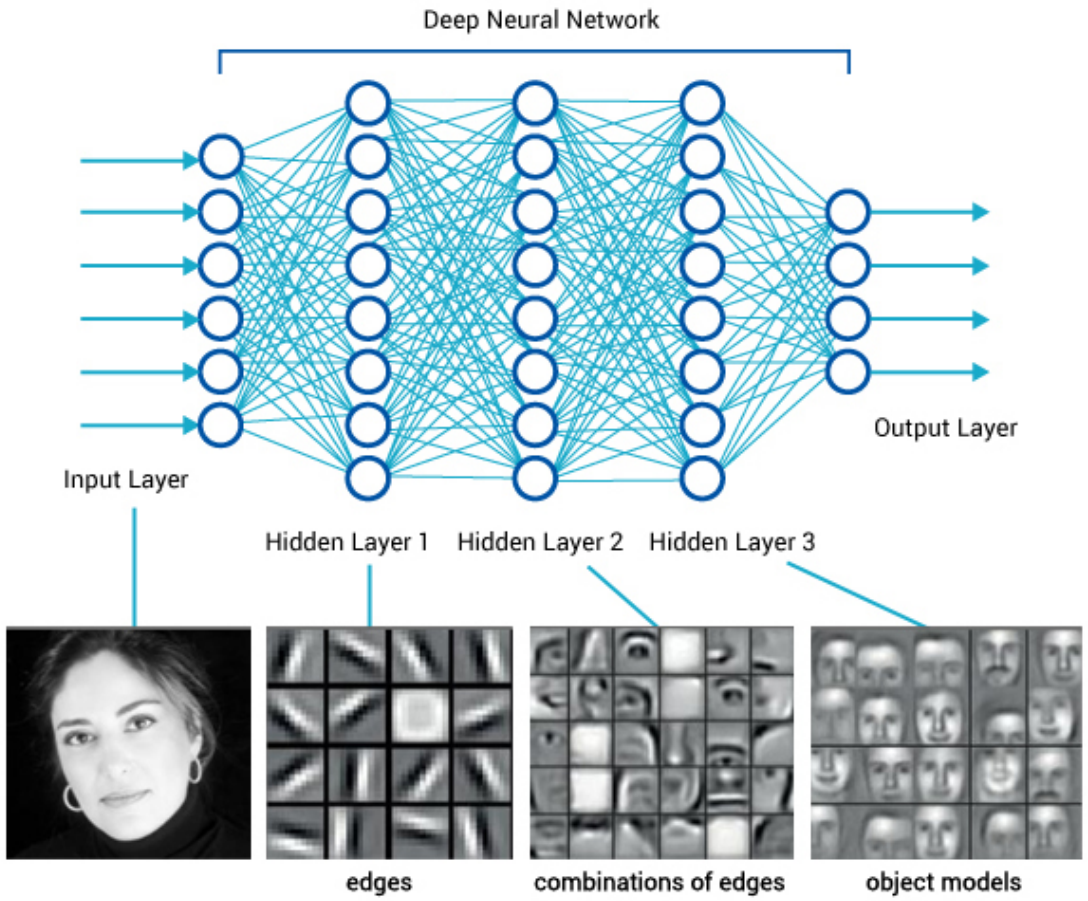
딥러닝 원리

What We See

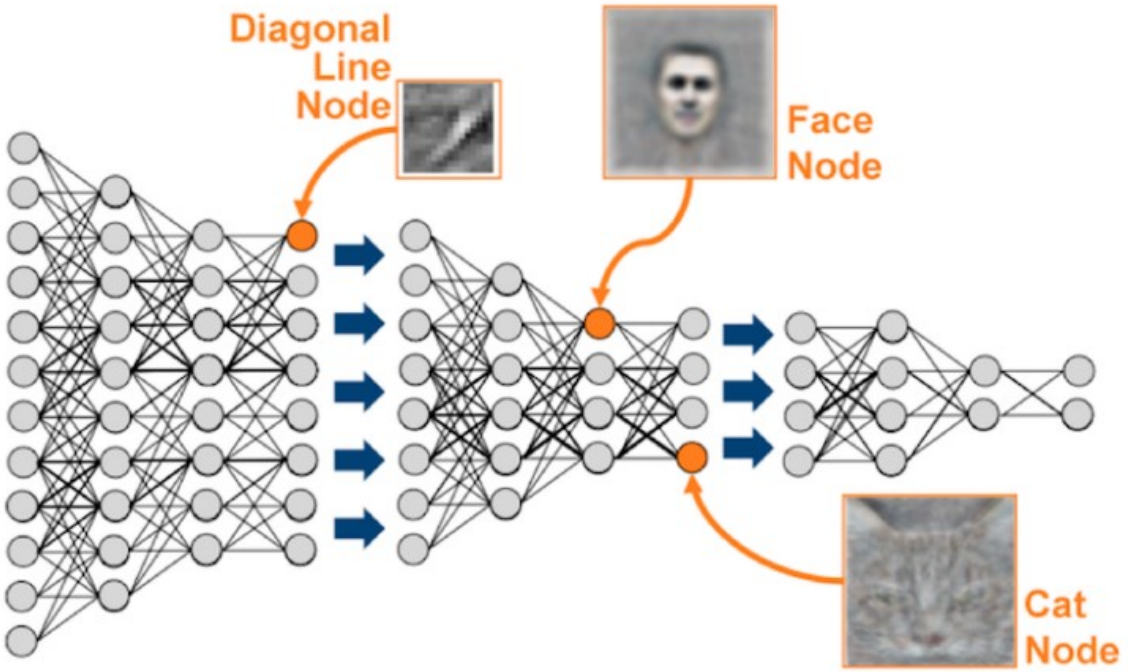
```
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
52 70 95 23 04 60 11 42 69 24 68 56 01 32 56 71 37 02 36 91
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 67 48
```

What Computers See

딥러닝 원리


Deep Neural Network
Input Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Output Layer
edges
combinations of edges
object models

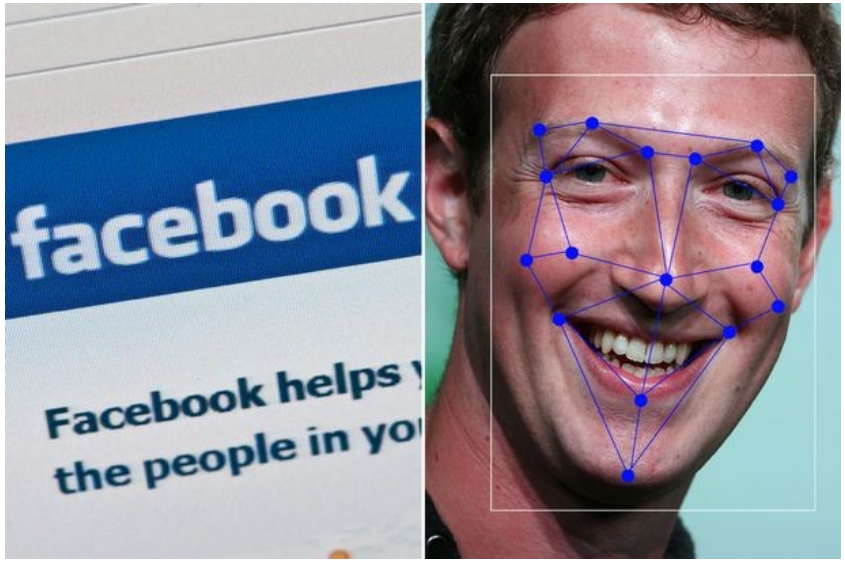
딥러닝 원리 – 컴퓨터가 정보를 추상화 하는 방법



Contents

딥러닝 사례 – ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

대용량의 이미지셋을 가지고 이미지 인식 알고리즘의 성능을 평가하는 대회

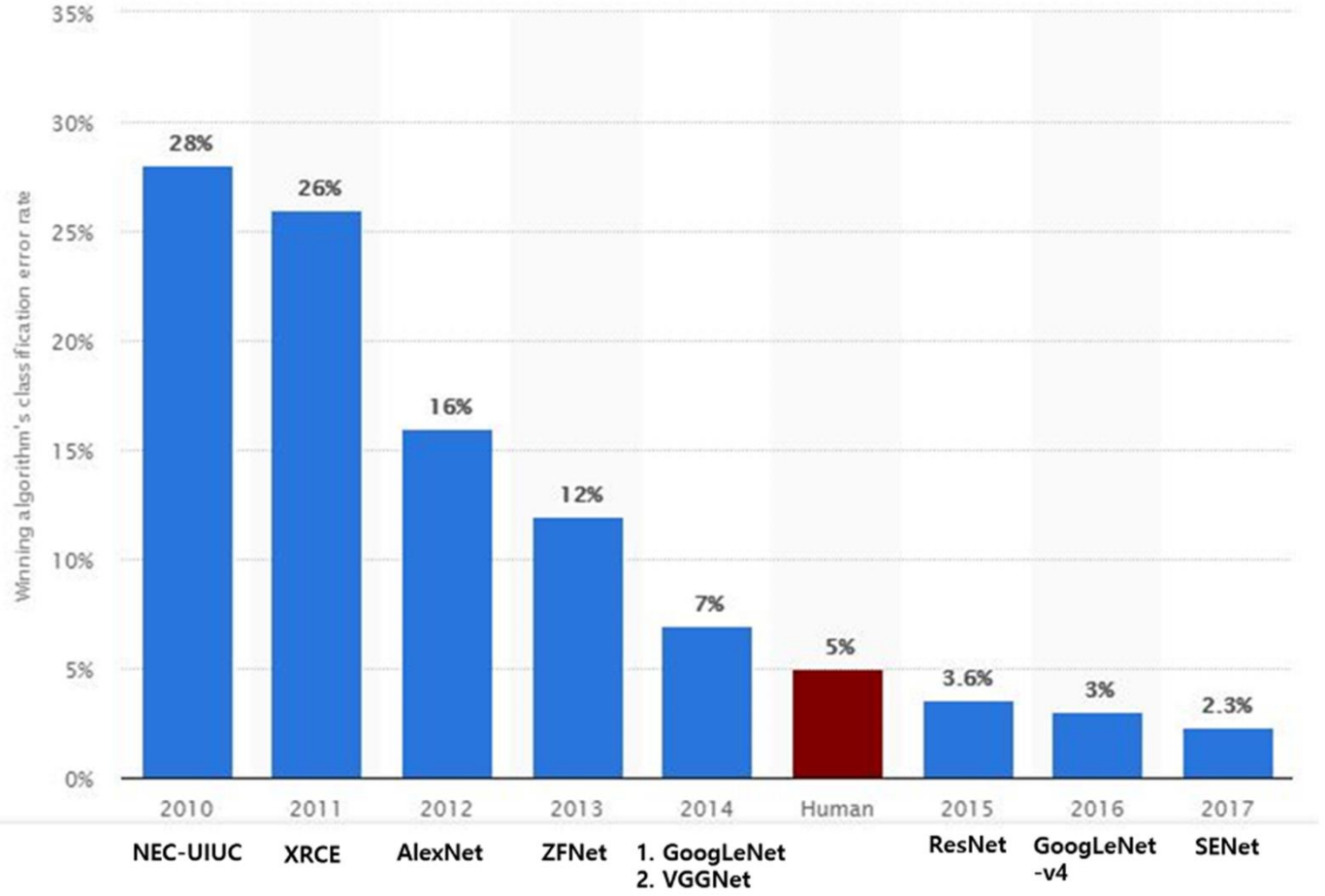


Face recognition

치와와 vs. 머핀

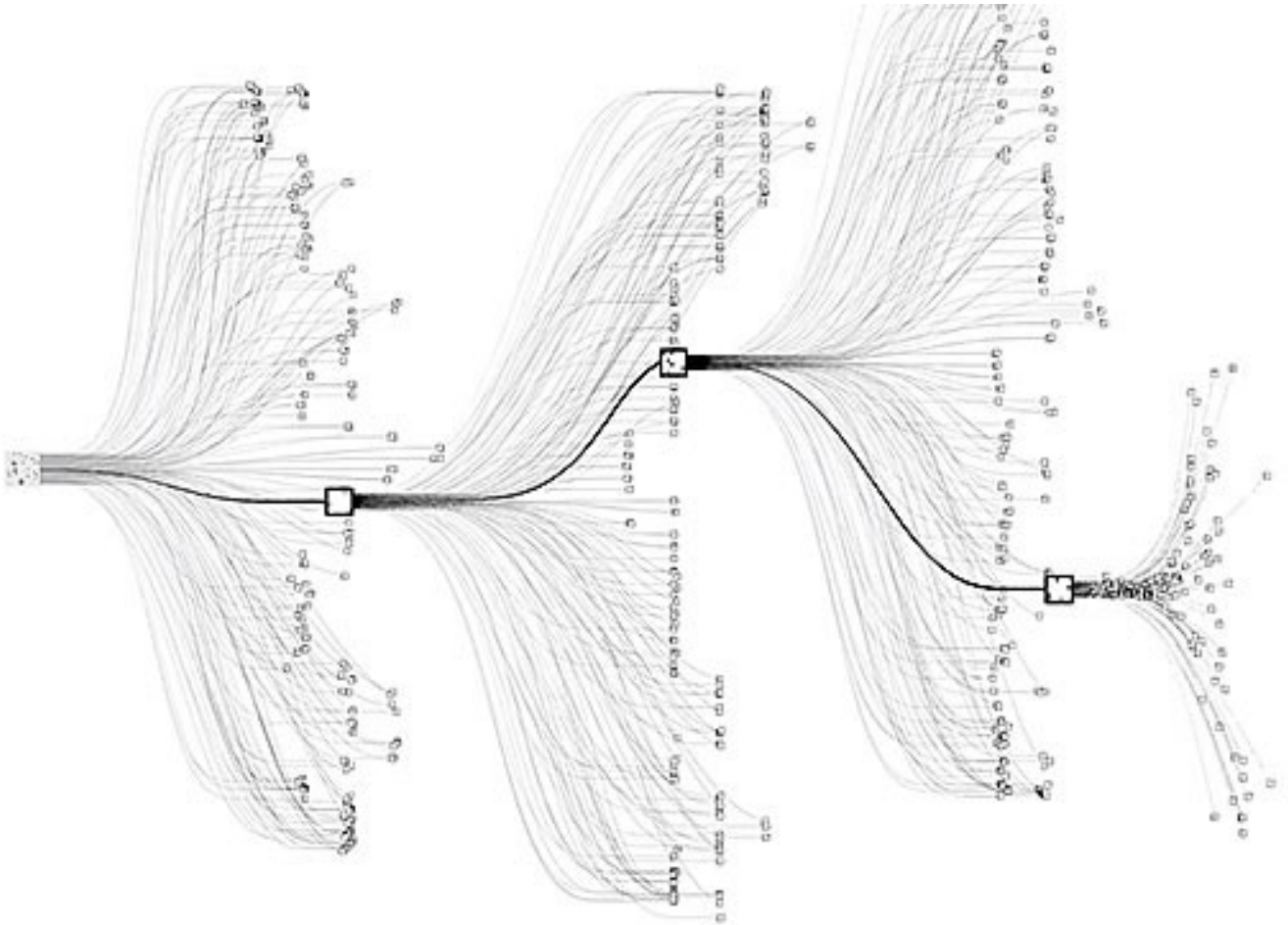
딥러닝 사례 – ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

대용량의 이미지셋을 가지고 이미지 인식 알고리즘의 성능을 평가하는 대회



딥러닝 사례 – 알파고 (1/3)

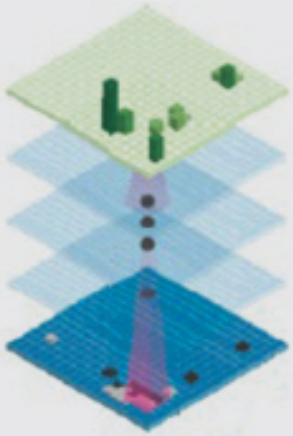
딥러닝 사례 – 알파고 (2/3)



딥러닝 사례 – 알파고 (3/3)

알파고의 정책망과 가치망 구조

정책망



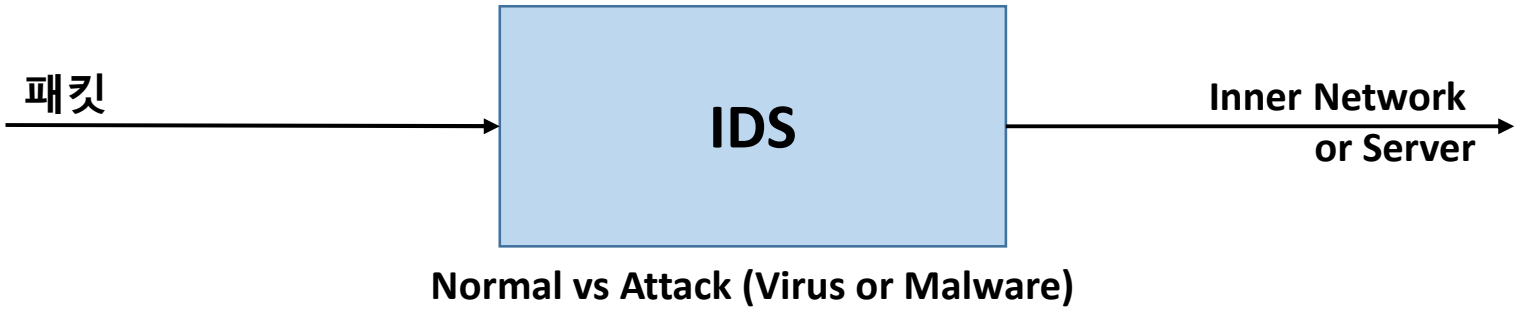
입력된 기보를 바탕으로
특정 상황에 전문가들이 많이 두는
수로 탐색의 폭을 좁혀주는 것

가치망

걸러진 수 가운데 현재 대국
상황에서 승산이 높은 수를 예측.
형세를 판단해주는 것

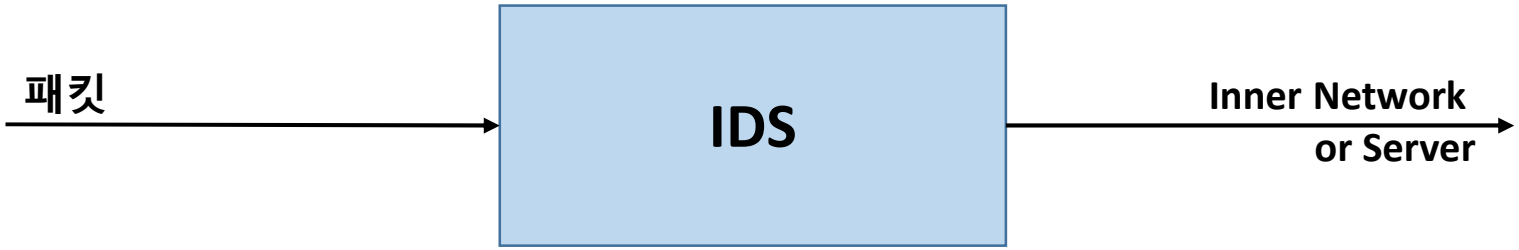
Contents

딥러닝과 보안 – Intelligent IDS (Intrusion Defense System) 예제



How?

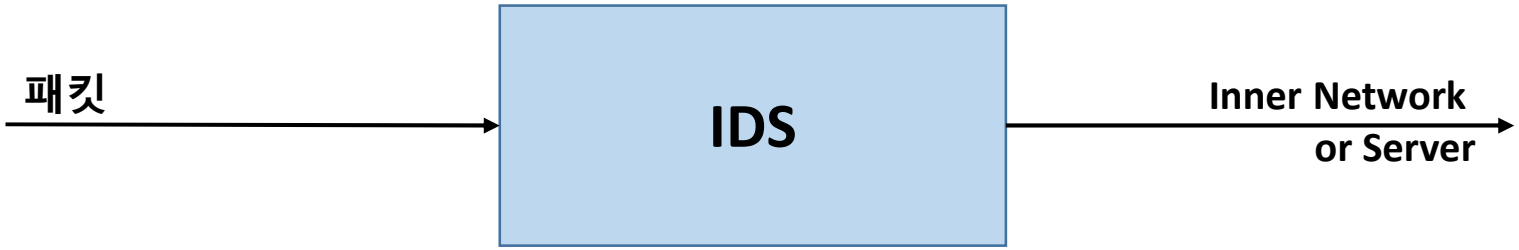
딥러닝과 보안 – Intelligent IDS (Intrusion Defense System) 예제

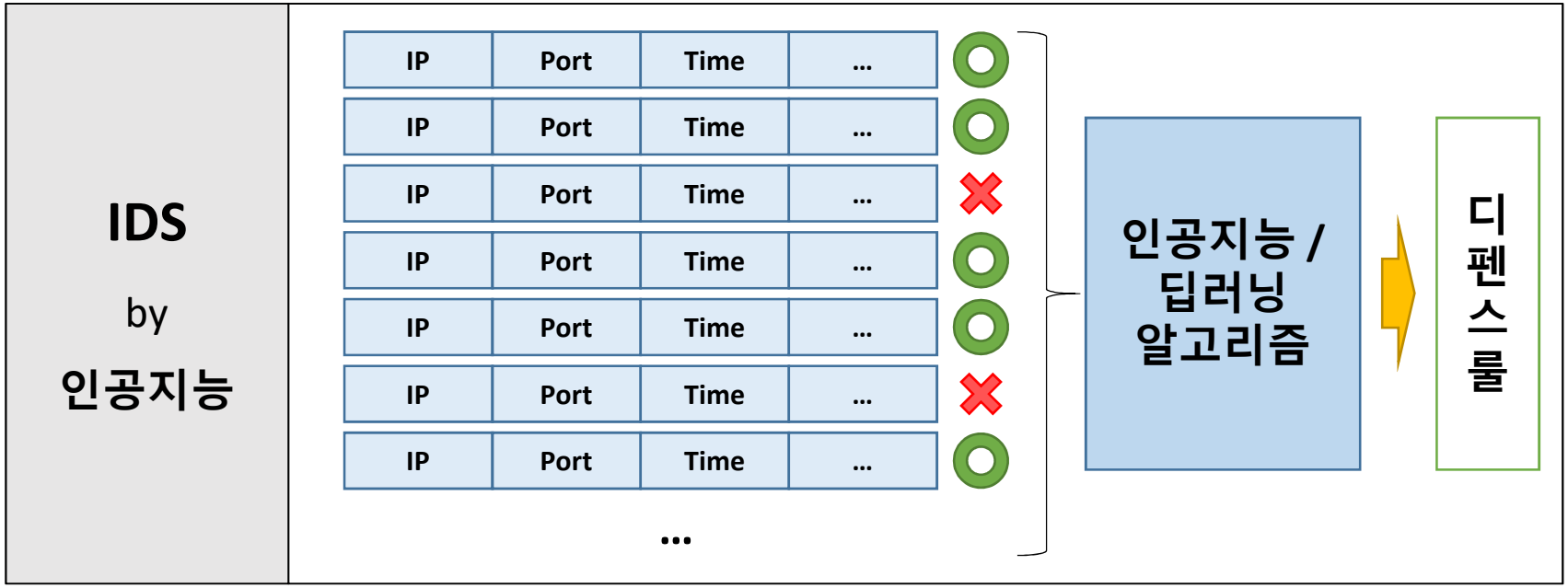


Normal vs Attack (Virus or Malware)

IDS by **Rule-base Engine**	**IDS : 해킹 공격은 어떤 것인가요?** **사람 : AAA.BBB.CCC.DDD 대역에서 들어오는 구만.~** **IDS : 그럼, AAA.BBB.CCC.DDD 대역 다 막을께요.** **사람 : 그건 아니고, Port 번호가 00인 것이 공격인 거 같아.** **IDS : 그럼 AAA.BBB.CCC.DDD 대역에 00 포트 번호는 다 막을께요.** **사람 : 아?! 잠깐만… 지금은 또 정상적인 접근인 거 같은데?** **IDS : 흠….**

딥러닝과 보안 – Intelligent IDS (Intrusion Defense System) 예제

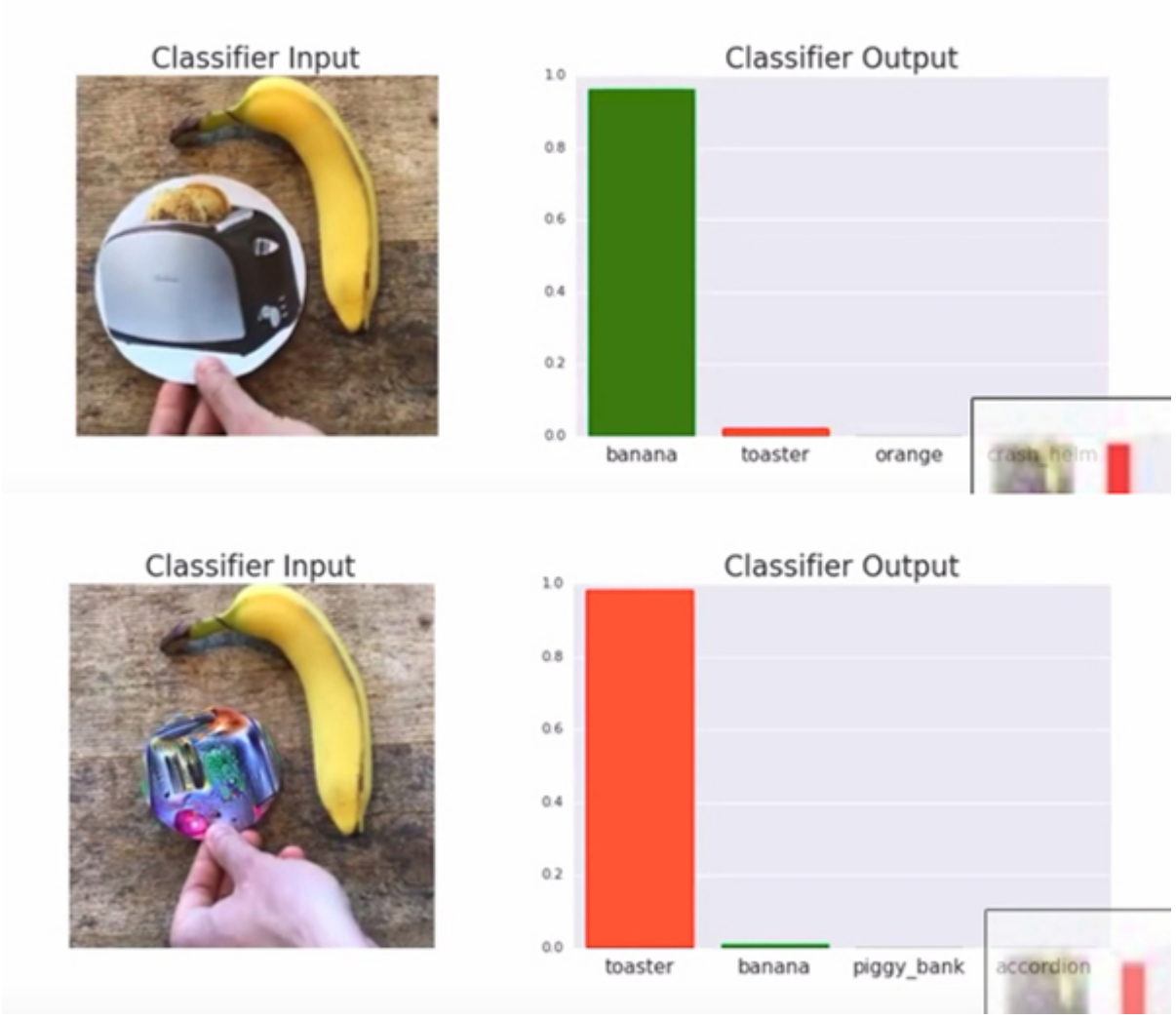




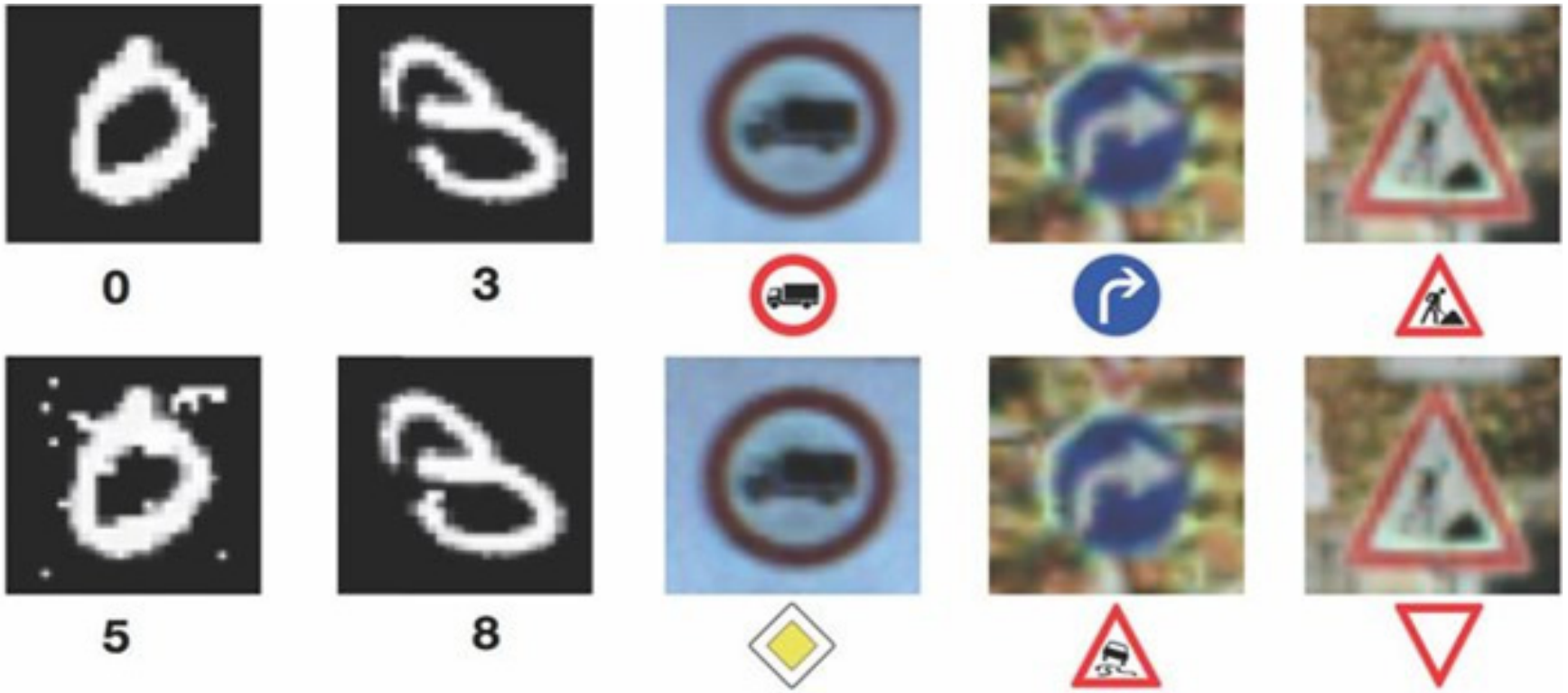
Contents

새로운 보안 위협

새로운 보안 위협


Classifier Input
Classifier Output
1.0
0.8
0.6
0.4
0.2
0.0
banana
toaster
orange
Classifier Input
Classifier Output
1.0
0.8
0.6
0.4
0.2
0.0
toaster
banana
piggy_bank
accordion

새로운 보안 위협



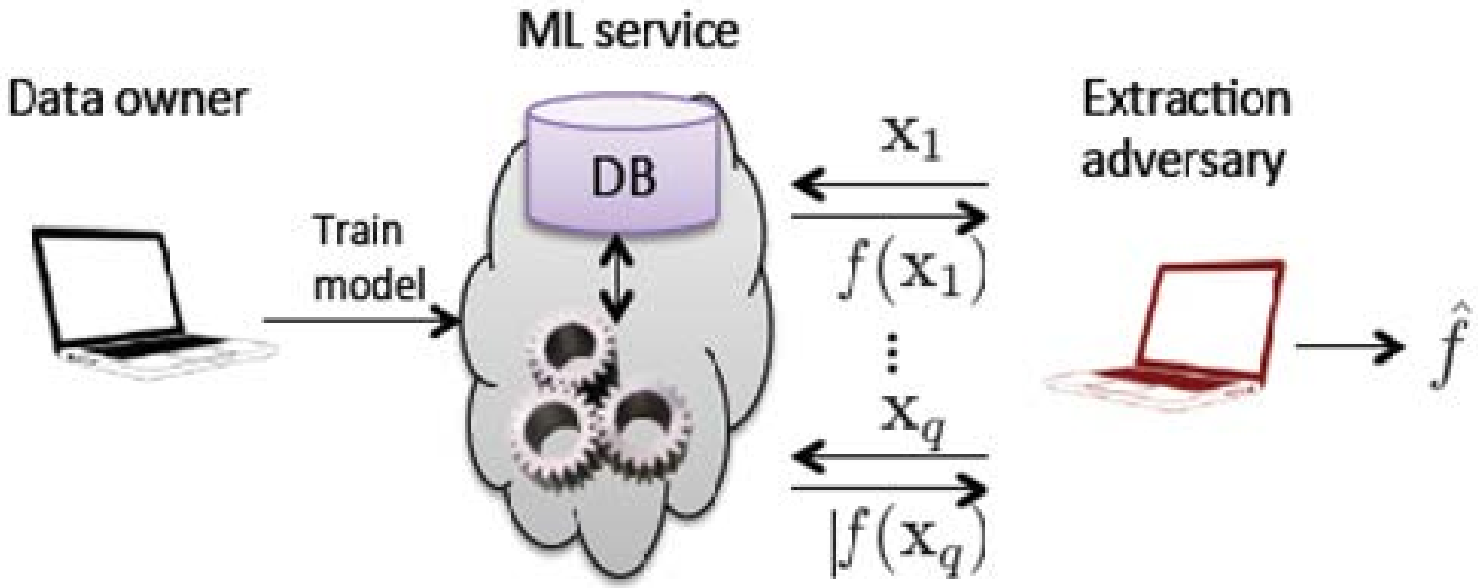
새로운 보안 위협

자율 주행 자동차의 Evasion Attack 예제

새로운 보안 위협

중국의 얼굴인식 결제 예제

새로운 보안 위협



Model Extraction Attack

참고문헌

- 전통적 프로그래밍 vs 인공지능 : https://www.datasciencecentral.com/profiles/blogs/traditional-programming-versus-machine-learning-in-one-picture
- 힐베르트 프로그램 & 괴델의 불완전성의 정리 : http://www.snunews.com/news/articleView.html?idxno=11883
- 튜링 머신1 : http://www.aistudy.co.kr/math/mechanism_math.htm
- 튜링 머신2 : https://norman3.github.io/papers/docs/neural_turing_machine.html
- 인공지능과 우리뇌에서, 구별하기와 표상하기 : http://scienceon.hani.co.kr/406294
- 인공지능은 의료를 어떻게 발전 시킬것인가? : http://www.yoonsupchoi.com/2017/08/08/ai-medicine-4/
- 고양이의 뇌 실험을 통한 Convolution 확인 : https://distillery.com/blog/implementing-human-brain-exploring-potential-convolutional-neural-networks/

참고문헌

- Deep Instinct : https://www.youtube.com/watch?v=Zmb_r981Vj8
- 박소희, 최대선 . 인공지능 보안 이슈. 한국정보보호학회. 2017. 27(3). 27-31
- 김종현. 인공지능 기반 금융권 보안관제 동향 및 향후과제. 전자금융과 금융보안. 2017. 제8호. 40-63
- ILSVRC 대회 (이미지넷 이미지 인식 대회) 역대 우승 알고리즘들 : https://bskyvision.com/425
- 무작위로 모든 경우의 수 따진다? '많이 두는 수' '승률 높은 수' 추려 : http://news.khan.co.kr/kh_news/khan_art_view.html?art_id=201603112154325
- 인공지능 (AI) 이미지 인식기술 무력화 '패치' 발견 : http://www.itbiznews.com/news/articleView.html?idxno=8430
- 얼굴만으로 계산 척척, 중국 안면인식 결제 시대 활짝 : http://www.newspim.com/news/view/20170906000099

감사합니다

최승우
sw.choi@kaist.ac.kr