의료 데이터를 활용한 데이터 분석

의료 영상 이미지를 활용한 딥러닝 실습

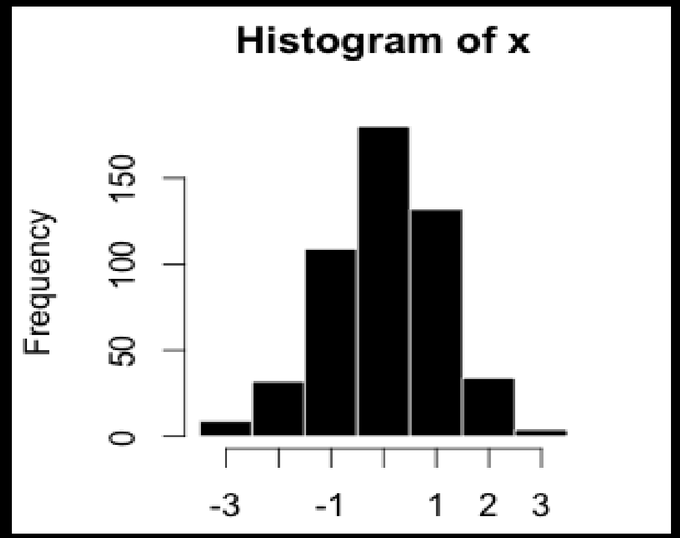
목차 의료 데이터를 활용한 데이터 분석

1. 히스토그램

히스토그램이란

- 수치 데이터의 분포를 이미지로 표현한 것
- 막대 그래프와 유사하게 연속적인 숫자 범위로 그룹화되는 많은 데이터를 빈으로 묶어 시각화

Bin/Interval	Count/Frequency
−3.5 to −2.51	9
−2.5 to −1.51	32
−1.5 to −0.51	109
−0.5 to 0.49	180
0.5 to 1.49	132
1.5 to 2.49	34
2.5 to 3.49	4

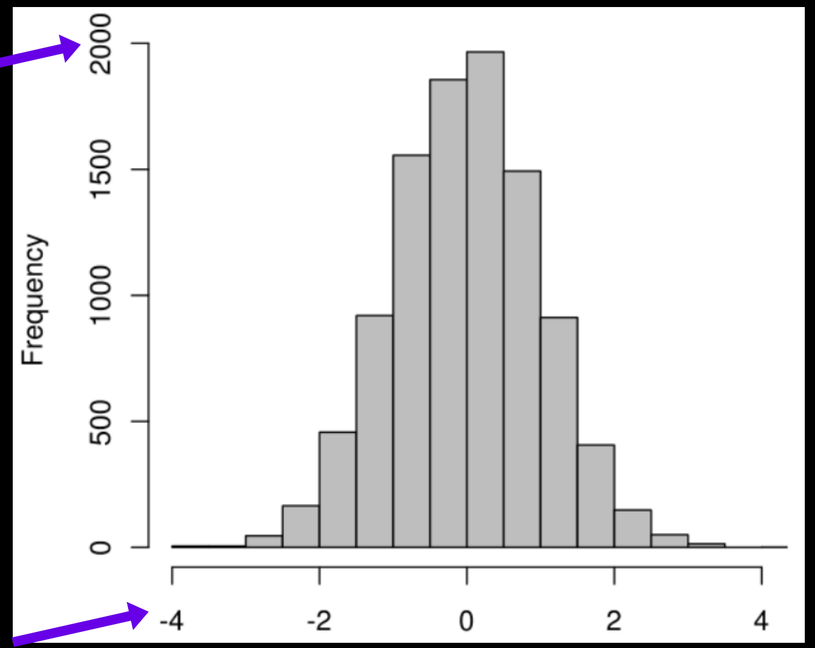


출처: https://en.wikipedia.org/wiki/Histogram

히스토그램이란

- 일반적으로 y축은 관측치를 나타냄
- X축은 빈(혹은 인터벌)이라고 하는 연속적인 숫자 범주를 나타냄

관측치



빈(bin)

출처: https://en.wikipedia.org/wiki/Histogram

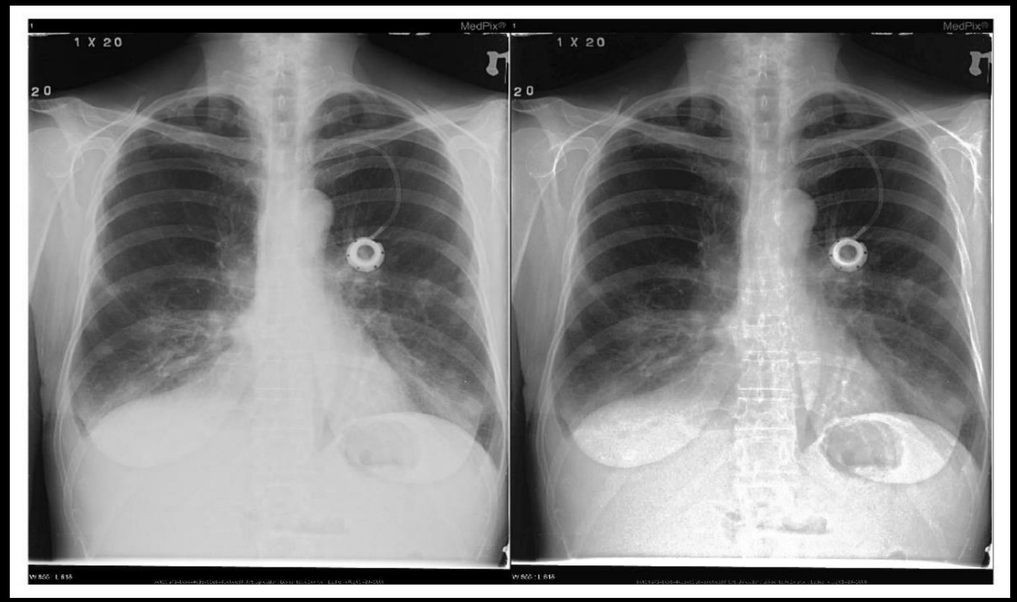
이미지의 대비 향상(Contrast Enhancement)

- 시각적인 측면에서 막대그래프와 히스토그램의 차이는
  막대그래프와 달리 모든 막대가 연속적으로 붙어있다는 것

막대 그래프	히스토그램
X축은 어떤 값도 표현 가능	X축은 연속된 숫자 데이터만 표현 가능
일반적으로 연속된 두 막대 사이에 일정한 간격이 있음	두 개의 연속된 두 막대사이에는 간격이 없음

이미지의 대비 향상(Contrast Enhancement)

- 이미지 대비 향상은 널리 사용되는 이미지 전처리 기술 중 하나
- 흐릿한 이미지를 선명하게 만들어 모델의 성능을 향상



Kumari et al., Contrast Enhancement of Medical Radiography Images Using Edge Preserving Filters, ICBSII 2018.

이미지의 대비 향상(Contrast Enhancement)

- 의료 이미지의 품질 저하 요인
  - 데이터 수집 과정에서 발생하는 불가피한 노이즈
  - 고르지 않은 조명과 같은 여러 요인으로 인해 발생하는 낮은 대비

원인	설명
잡음	실제 입력되지 않았지만 입력 되었다고 잘못 판단된 값
결측값	데이터가 입력되지 않았지만 입력되었다고 잘못 판단한 값
이상값	데이터의 정상적인 범위에서 많이 벗어난 아주 크거나 작은 값
불일관성	다양한 형태/형식의 데이터

이미지의 대비 향상(Contrast Enhancement)

- 다양한 이미지의 대비 향상 기술들이 존재함
  1. 선형 대비 향상
  2. 비선형 대비 향상 (e.g. 감마 보정)
  3. 히스토그램 대비 향상 (e.g. 히스토그램 균등화)
  4. ...

-

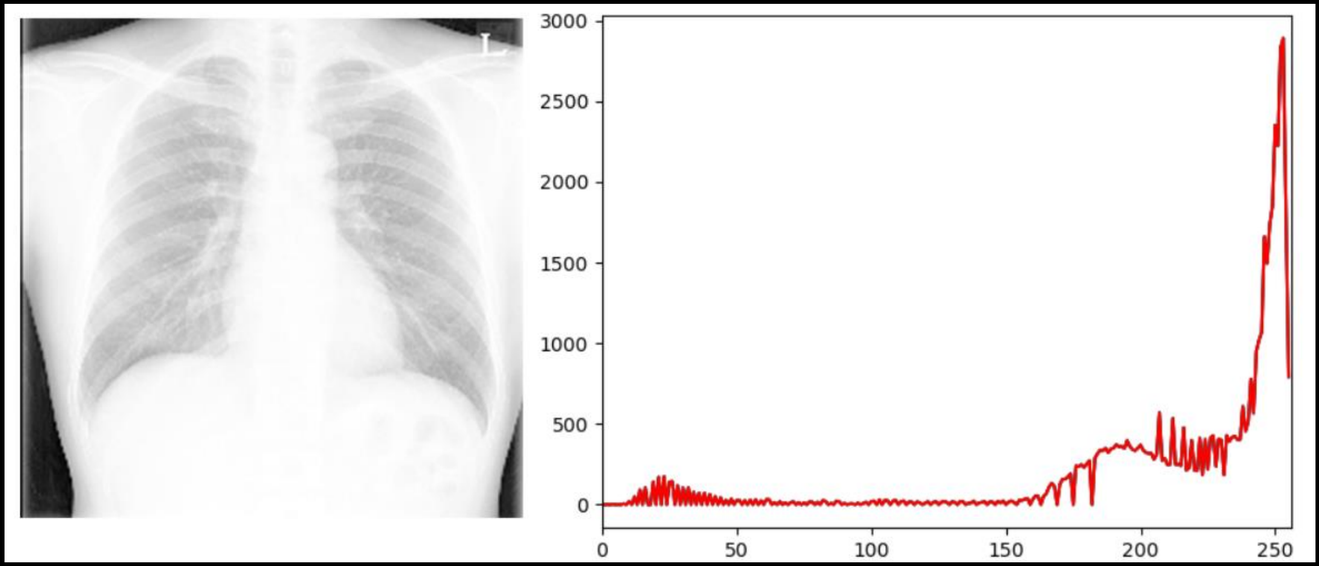
이미지의 대비 향상(Contrast Enhancement)

- 다양한 이미지의 대비 향상 기술들이 존재함
  1. 선형 대비 향상
  2. 비선형 대비 향상 (e.g. 감마 보정)
  3. 히스토그램 대비 향상 (e.g. 히스토그램 균등화)
  4. ...

-

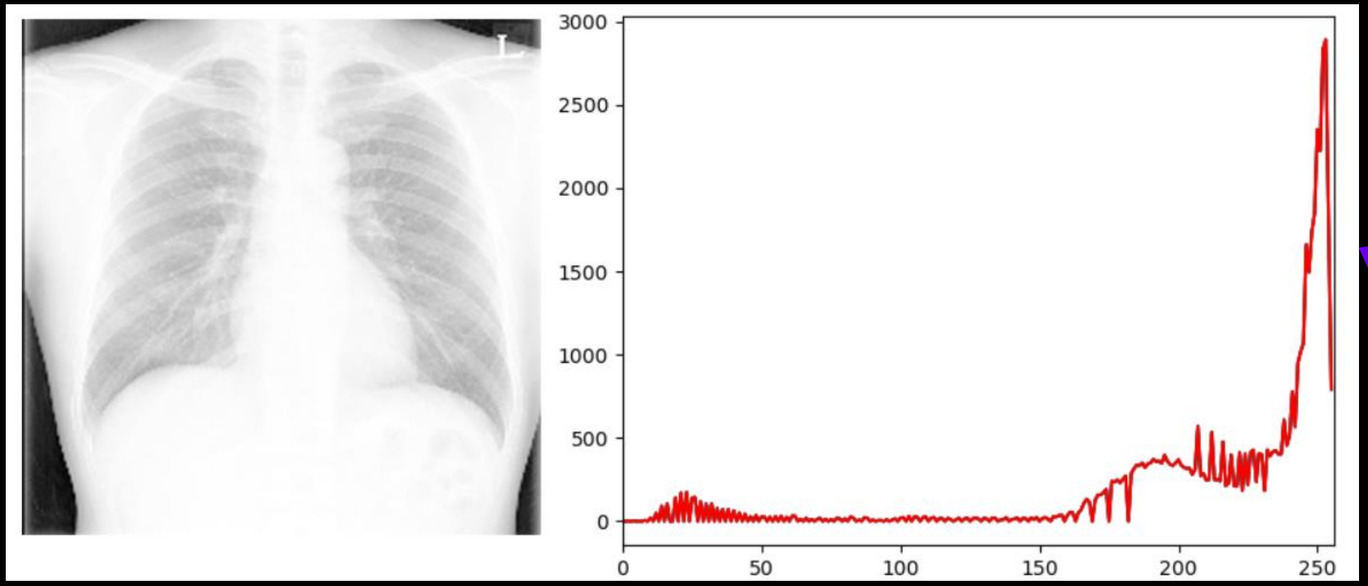
히스토그램 균등화(Histogram Equalization)

- 특정 그레이 레벨 사이에 몰려 있는 히스토그램 분포를 균등하게 늘리는 방법입니다.
- 즉, 변환 함수를 통해 히스토그램을 전체 그레이 레벨에 고르게 분포 시킵니다.



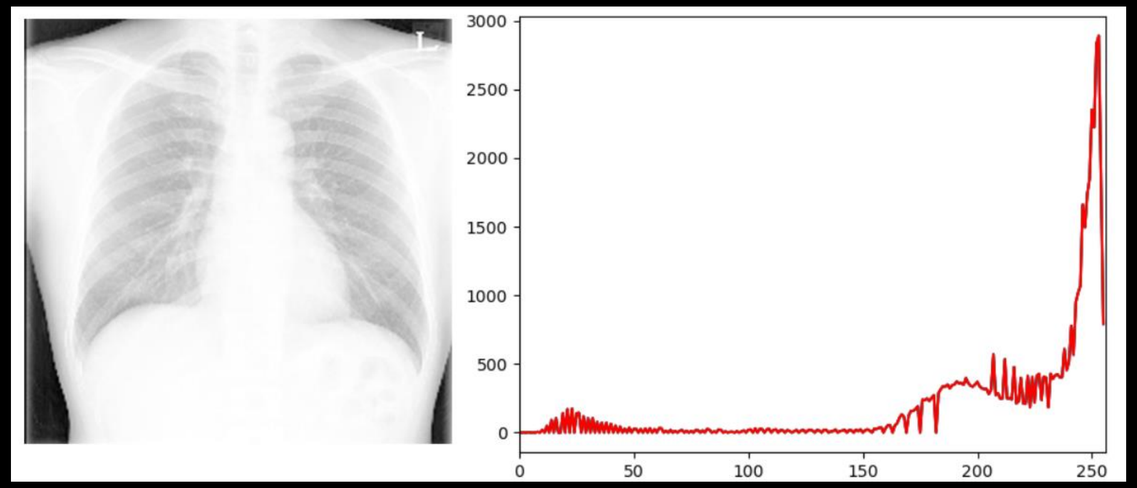
히스토그램 균등화(Histogram Equalization)

- 아래 그림은 과도한 조명으로 인해 거의 모든 픽셀들이 높은 명도 값을 가지는 왼쪽 기울어짐이 발생한 예시입니다.

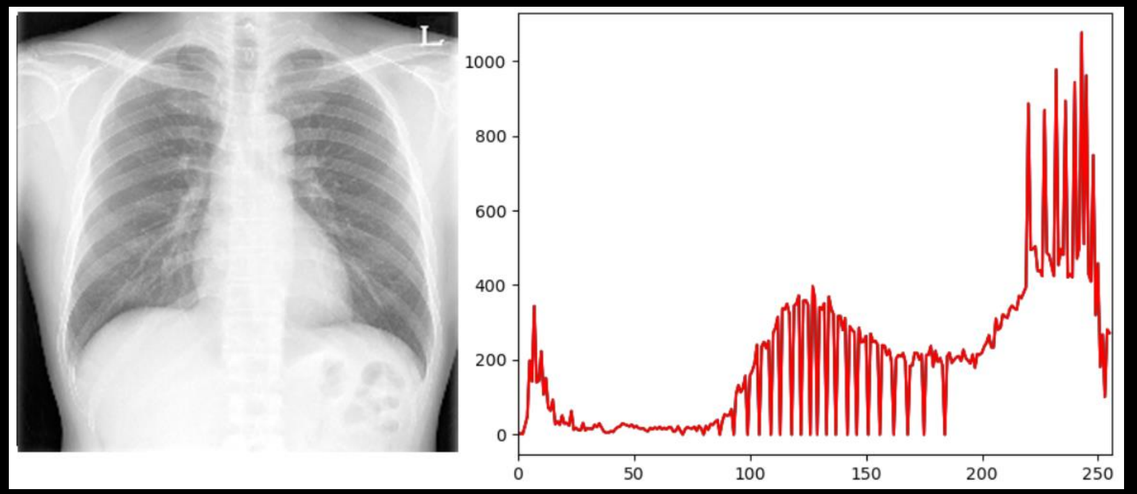


히스토그램 균등화(Histogram Equalization)

- 이때 히스토그램 균등화를 통해 전체적인 픽셀의 그레이 레벨의 분포를 균등하게 만들어주면 이미지를 보다 선명하게 관찰할 수 있음
- 이러한 이미지 대비향상은 위성, 열화상 및 X선 이미지 같은 과학 이미지에서 매우 유용한 기술



균등화 전 데이터



균등화 후 데이터

2. VGG16 이해하기

이미지 인식이란?

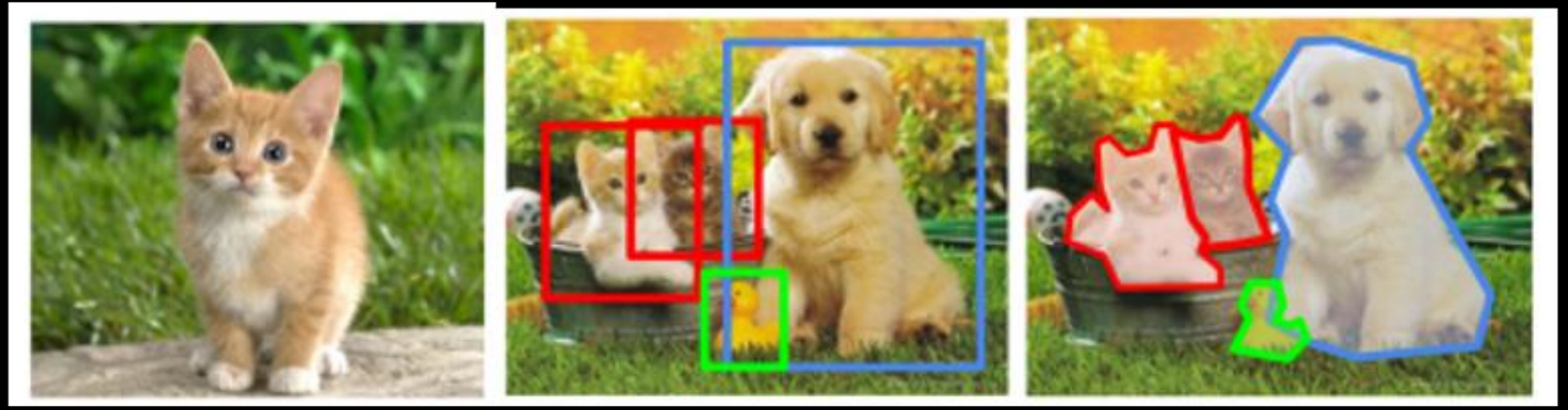
- 인간이 가진 시각적 능력을 컴퓨터 시스템으로 구현하는 작업
- 사람들이 매일 생성하는 수십억 장의 사진 및 동영상을 처리하기 위해 콘텐츠 레이블링, 이미지 검색 등 다양한 영역에서 이미지 인식이 활용되고 있습니다.

출처: https://online.stanford.edu/

이미지 인식이란?

- 대표적인 이미지 인식 작업으로는 이미지 분류(Image classification), 객체 탐지(Object detection), 이미지 분할(Semantic segmentation)이 있습니다.

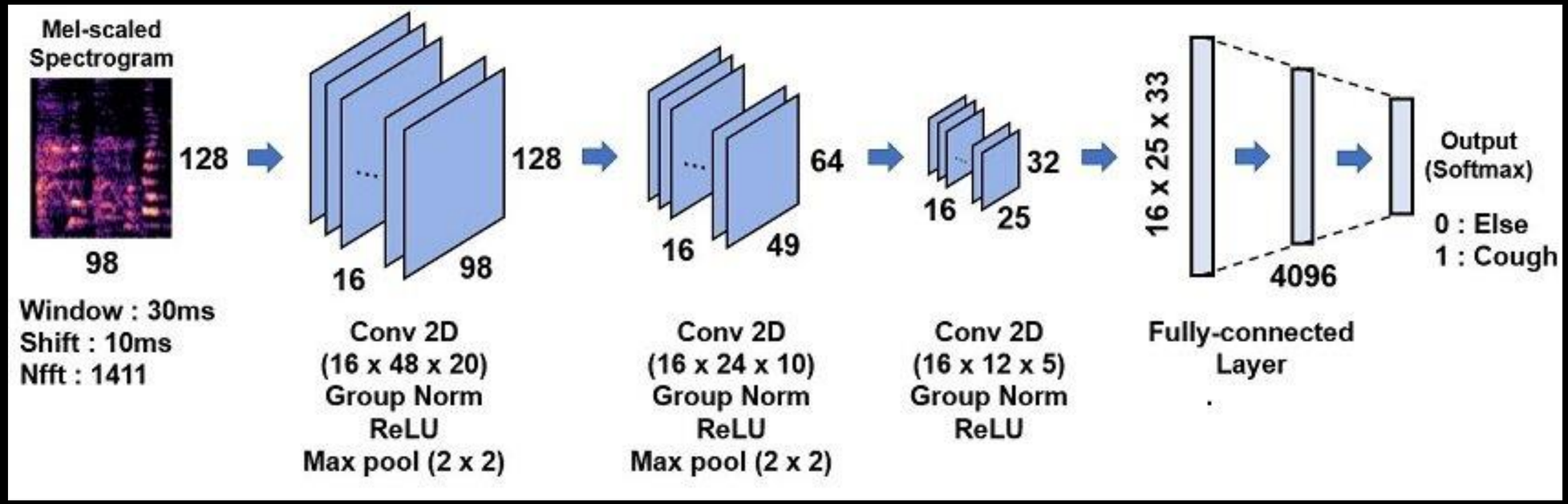
이미지 분류 | 객체 탐지 | 이미지 분할



출처: https://medium.com/

합성곱 신경망(Convolutional Neural Network, CNN)

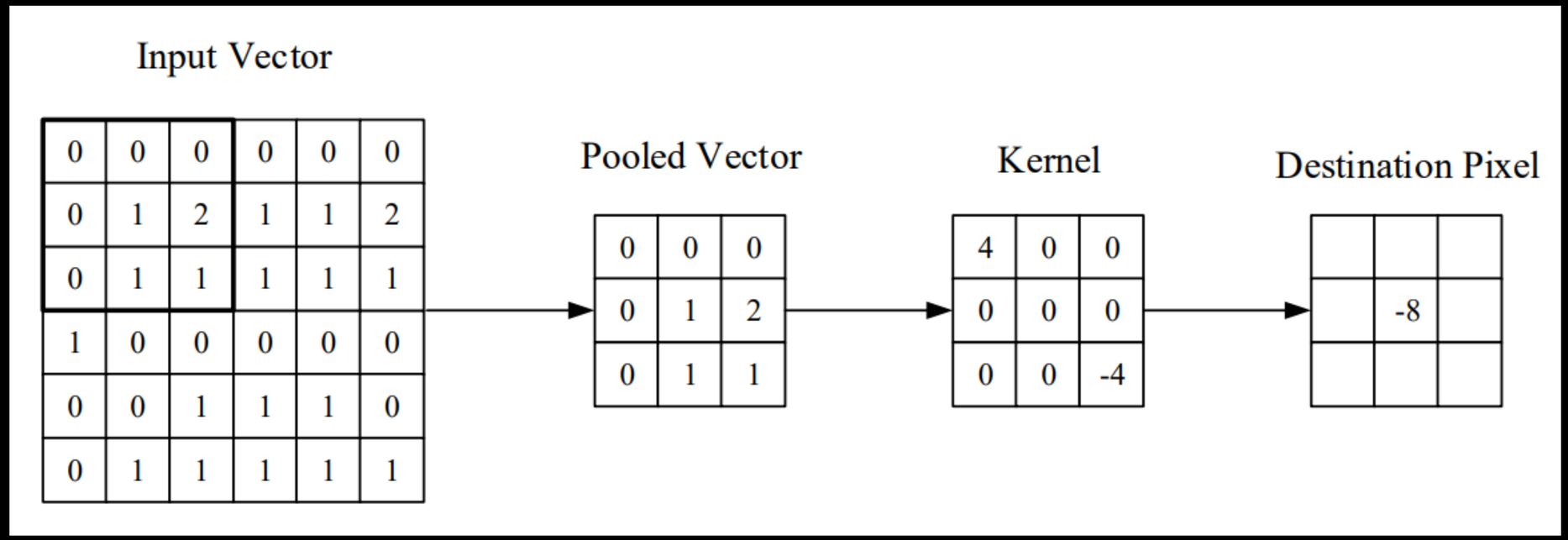
- 일반적으로 대부분의 이미지 인식모델은 여러 층의 합성곱 신경망으로 이루어져 있습니다.



출처: https://techxplore.com/

합성곱 신경망(Convolutional Neural Network, CNN)

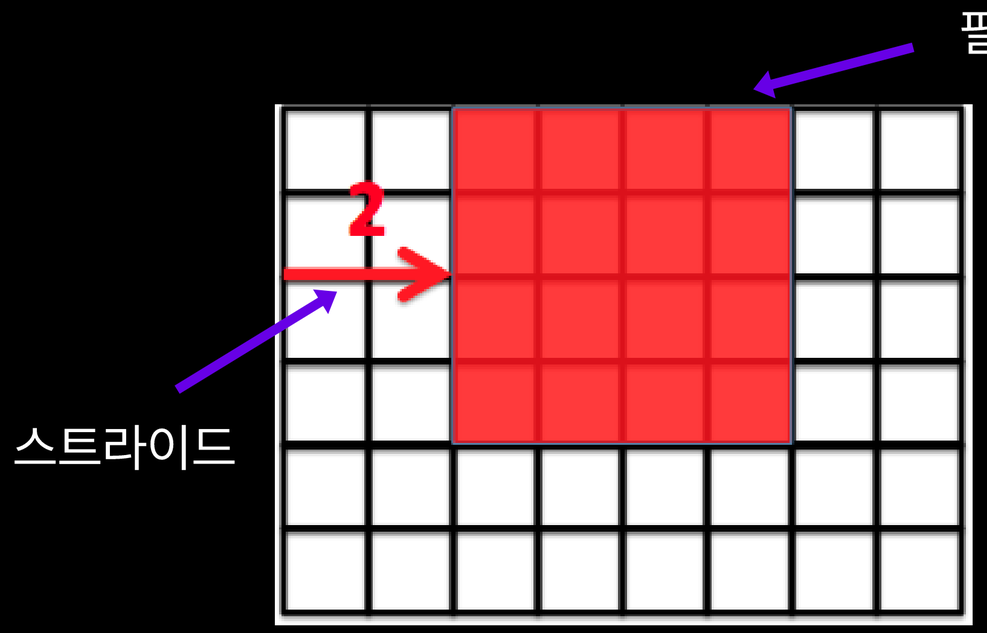
- 인간의 시신경 구조를 모방한 모델
- 합성곱 레이어를 사용하는 인공신경망 구조입니다.
- 이미지 인식 모델은 합성곱 신경망을 이용해 공간 계층 구조를 학습할 수 있습니다.



O'Shea et al., An Introduction to Convolutional Neural Networks, 2015

합성곱 신경망(Convolutional Neural Network, CNN)

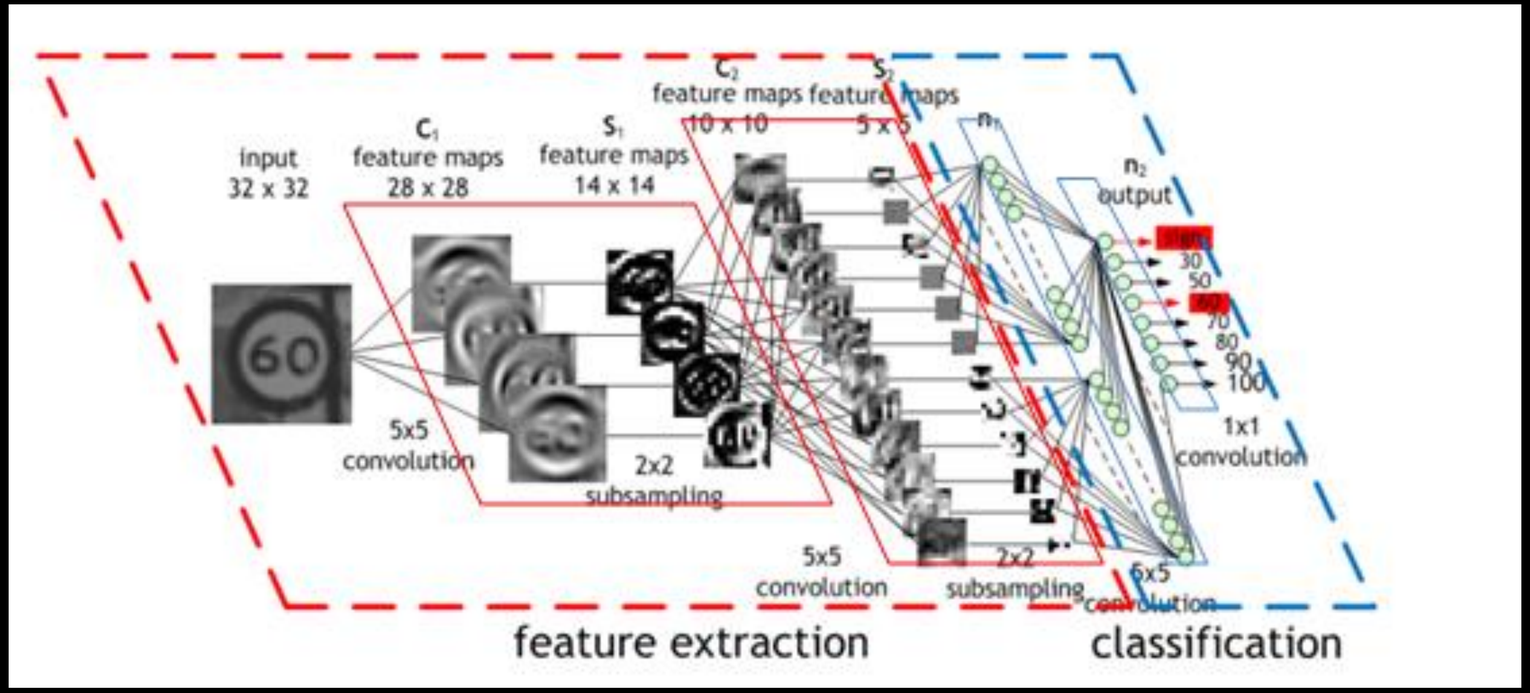
- 합성곱 신경망의 가장 중요한 단위인 합성곱 레이어는 필터, 스트라이드, 패딩으로 구성됩니다.
  - 필터(Filter): 이미지의 특징을 찾아내기 위한 파라미터
  - 스트라이드(Stride) 필터가 움직이는 지정된 간격
  - 패딩(Padding); 레이어의 출력 데이터가 줄어드는 것을 방지하기 위해 외곽에 특정 값을 채워 넣는 것
  -



출처: https://medium.com/

합성곱 신경망(Convolutional Neural Network, CNN)

- 합성곱 신경망은 입력 레이어, 출력 레이어 그리고 은닉 레이어(합성곱 레이어)로 구성됩니다.



출처: https://developer.nvidia.com/

VGG

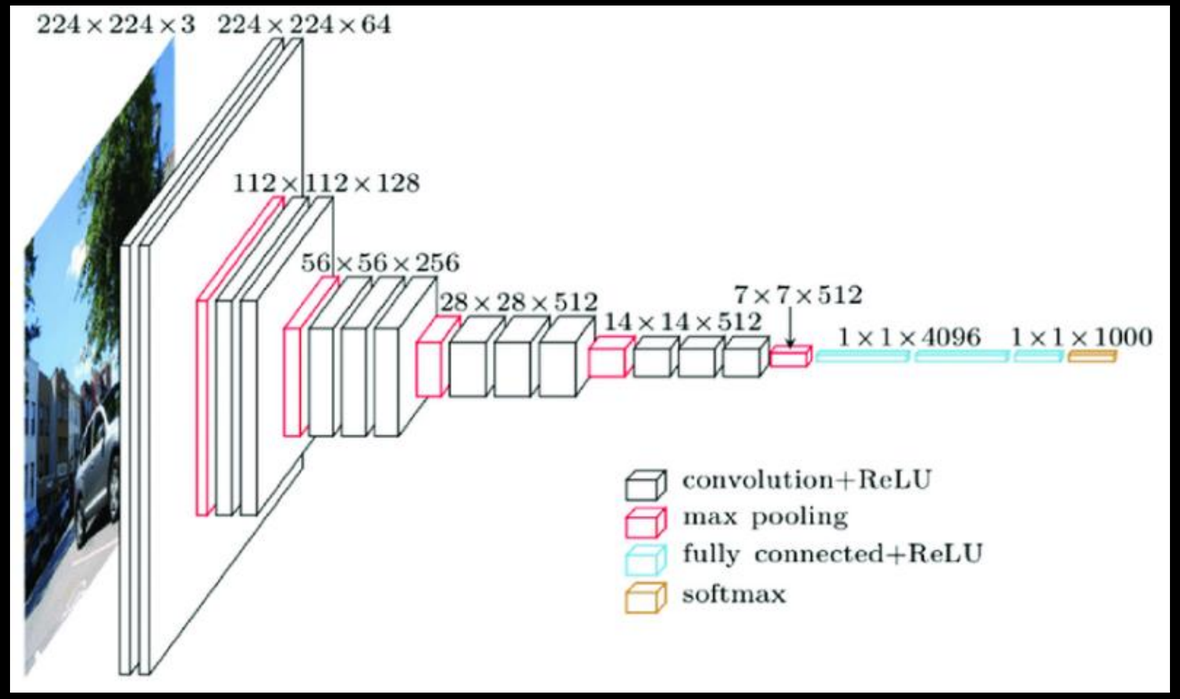
- Visual Geometry Group에서 개발한 모델
- VGG는 개발 당시 이미지 인식 경진대회 중 하나인 ILSVRC에서 최고 성능을 달성한 모델

Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

출처: https://medium.com/

- 강력한 이미지 인식 구조 중 하나로써 이미지 처리에서 필수적으로 알아야 하는 모델입니다.



VGG16

- VGG는 여러 층의 합성곱 신경망으로 이루어진 모델로 층의 개수에 따라 VGG16 혹은 VGG19로 명명됩니다.

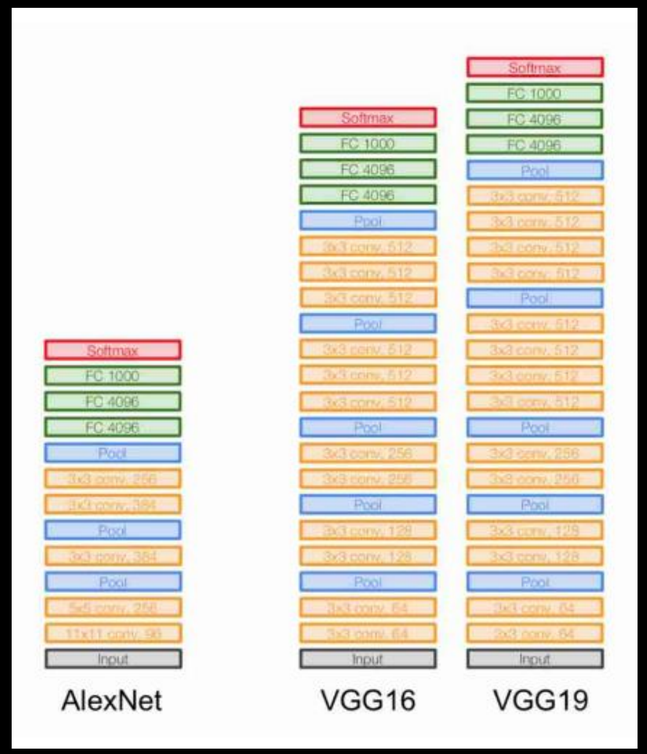


Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

3. Class Imbalance

Class Imbalance (데이터 불균형) 이란

- 어떤 데이터에서 각 클래스별로 갖고 있는 데이터의 양에 차이가 있을 경우, 클래스 불균형이 있다고 말합니다.

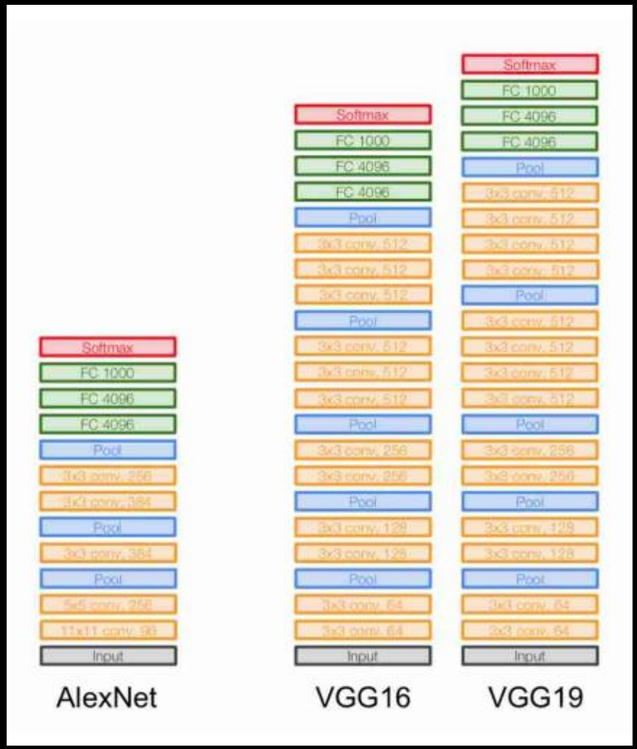
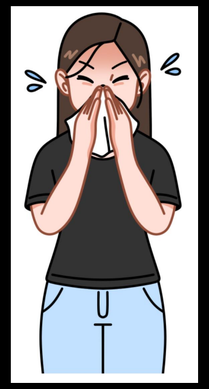


Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

Class Imbalance (데이터 불균형) 이란



<

1명

500명

아픈 사람보다 아프지 않은 사람이 훨씬 더 많습니다!

(클래스 불균형)

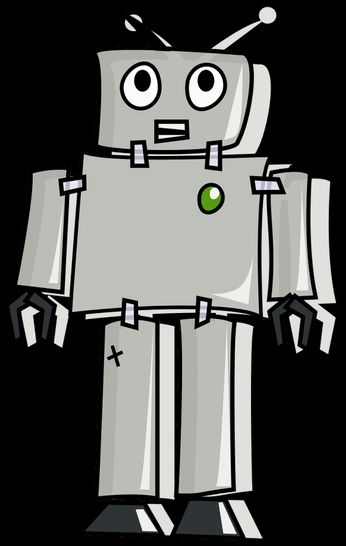
Class Imbalance (데이터 불균형) 이란

- 병이 있는지를 판단하는 문제에서는 모델이 '무조건 아프지 않다' 라고 판단하면 대부분의 환자를 맞출 수 있어, 진짜 환자도 '아프지 않다'고 잘못 분류될 확률이 높습니다.

1명

<

500명



' 무조건 아프지 않다고 판단하면

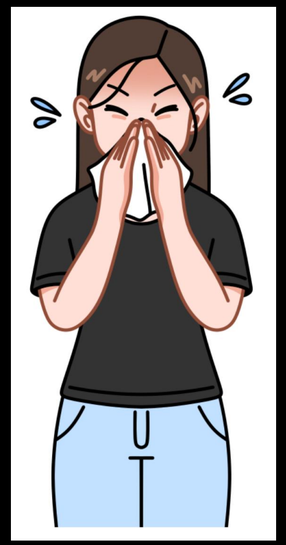
1명 빼고 다 맞추네?'

Weight Balance(가중치 조절)

- training set의 각 데이터에서 loss를 계산할 때

  특정 클래스의 데이터에 더 큰 loss 값을 갖도록 하는 방법입니다.

- 예를 들어 '환자'의 클래스에는 75%의 가중치를 두고, '정상'의 클래스에는 25%의 가중치를 두어 환자 클래스일 때의 loss에 더 민감하게 학습하도록 합니다.



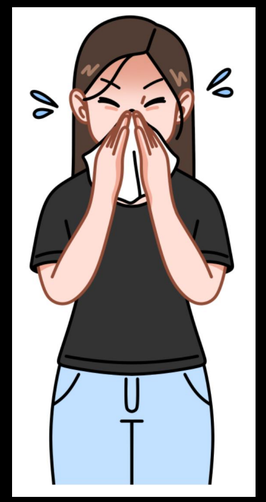
75%

25%

Weight Balance(가중치 조절)

- 이진 분류의 경우, 교차 엔트로피 식에서는 다음과 같이 α와 $1-\alpha$로 가중치를 줍니다.

$$loss = -\alpha t(\log(y) + (1-\alpha)(1-t)\log(1-y))$$

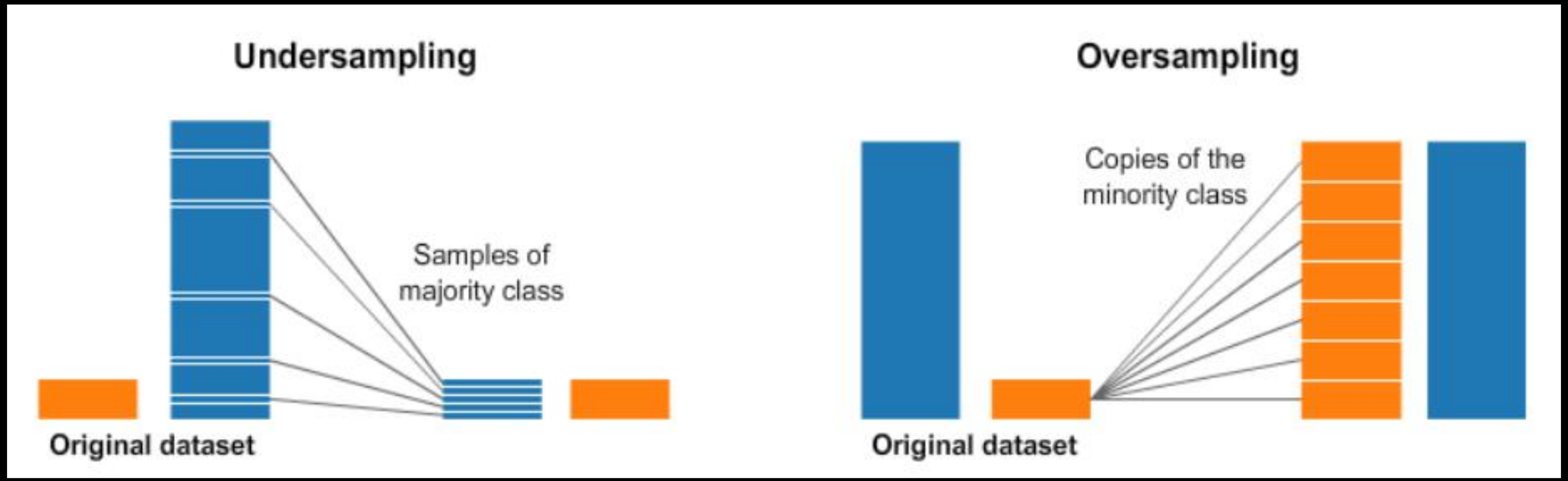


t = target (0/1)

y = prediction(0~1)

Oversampling & Undersampling

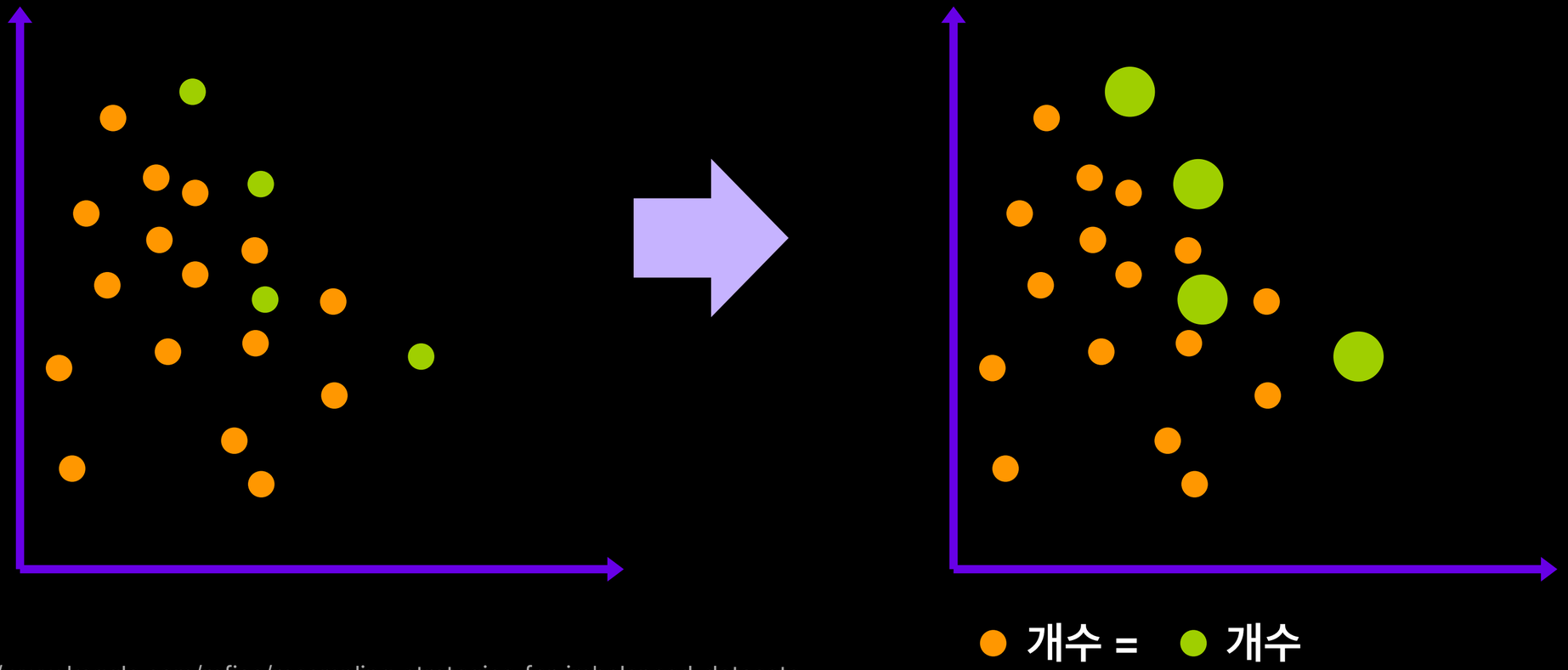
- Oversampling과 Undersampling은 학습 데이터 생성 과정에서 정상과 비정상 데이터 분포를 맞춰서 분류할 클래스마다 비슷한 수의 instance를 갖게 하는 방법입니다.



출처: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Oversampling

- 간단한 랜덤 오버샘플링 방법
- 기존에 존재하는 소수의 클래스 데이터를 복제하여 비율을 맞춰줍니다.



출처: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Undersampling

- 배치를 만들 때 클래스에서 같은 개수만큼 샘플링하여 학습을 진행하는 방법입니다.