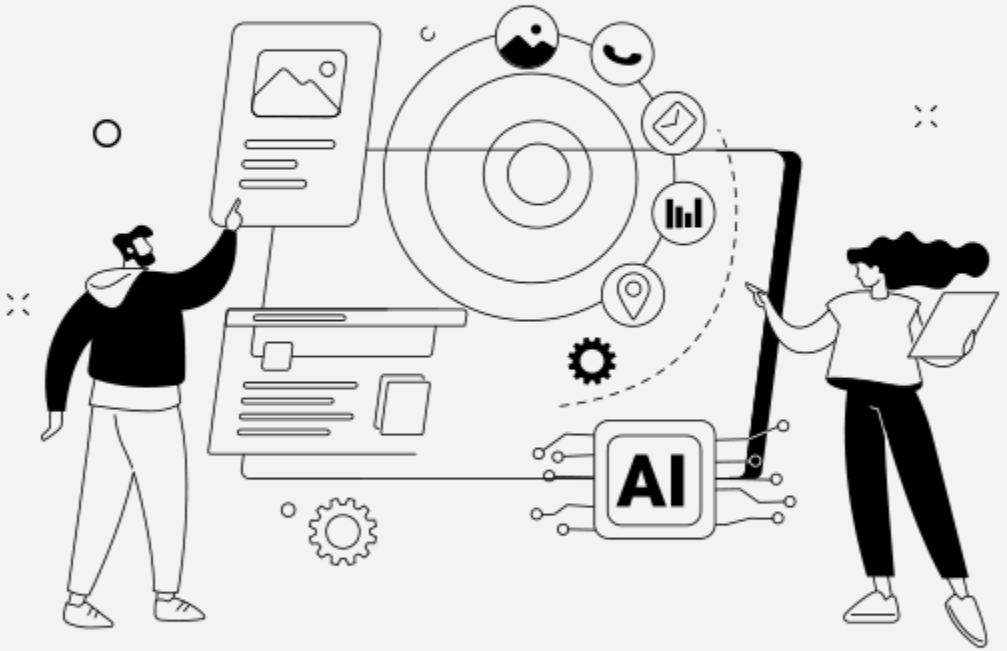
2022 데이터 크리에이터 캠프

Data Creator Camp



대학팀 0824

과학기술정보통신부 NIA 한국지능정보사회진흥원

팀 소개

연준

EDA

재우

실사 이미지 제거

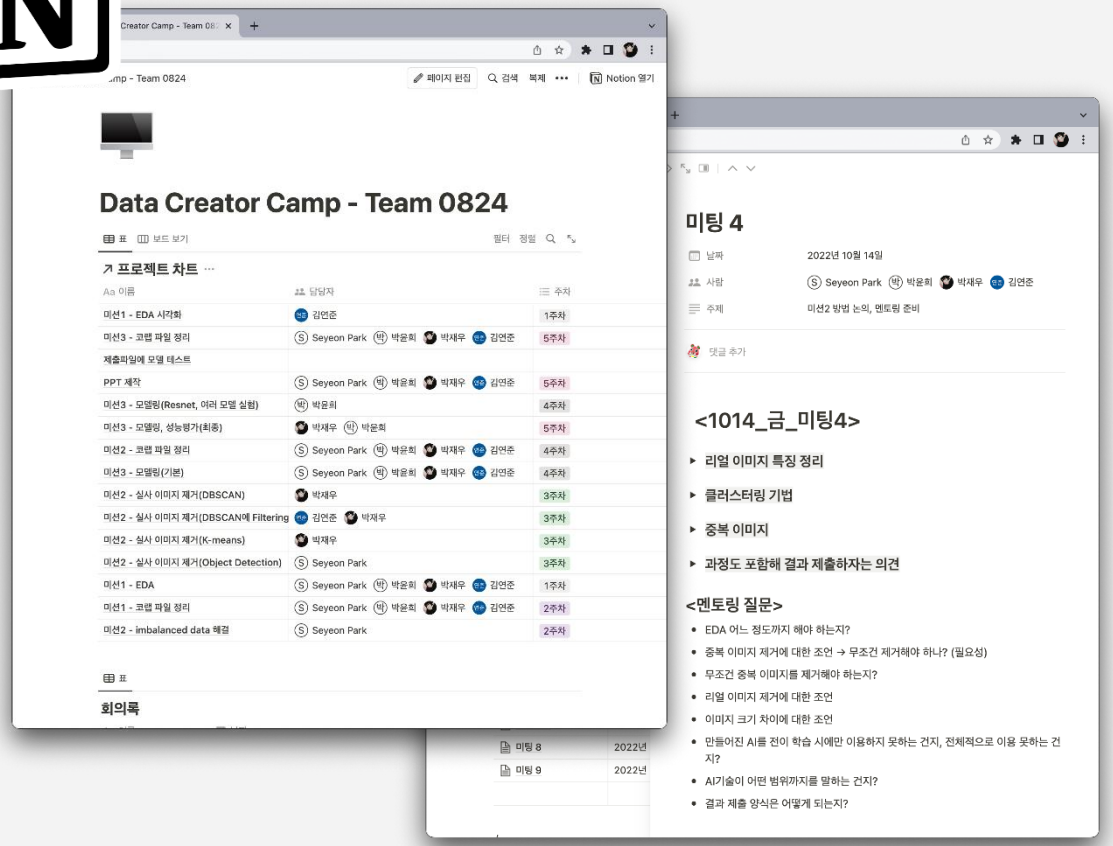
세연

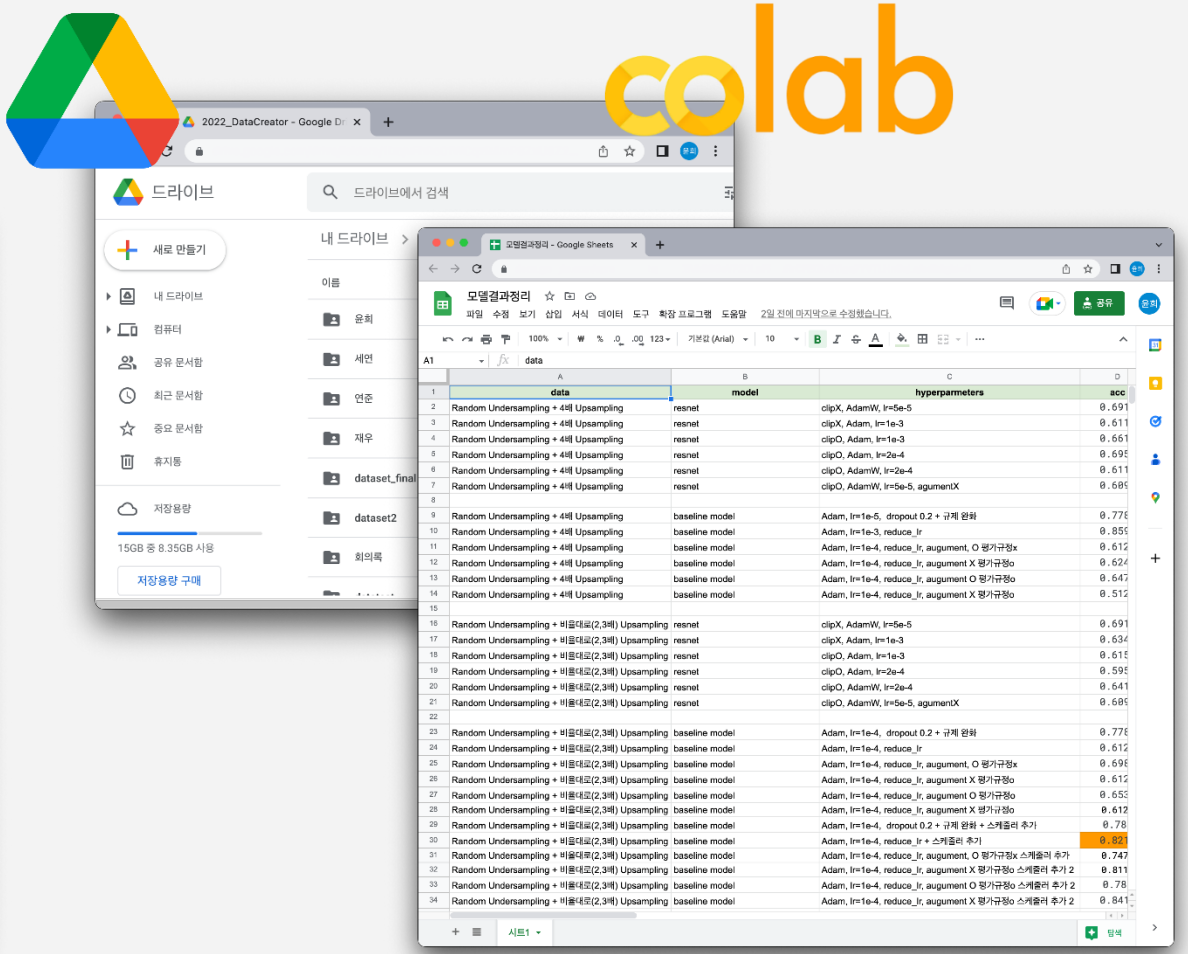
데이터 불균형

윤희

모델 구축

팀 협업 방법


Data Creator Camp - Team 0824
프로젝트 차트
회의록
미팅 4
2022년 10월 14일
Seyeon Park 박윤희 박재우 김연준
미션2 방법 논의, 멘토링 준비
<1014_금_미팅4>
리얼 이미지 특징 정리
클러스터링 기법
중복 이미지
과정도 포함해 결과 제출하자는 의견
<멘토링 질문>
EDA 어느 정도까지 해야 하는지?
중복 이미지 제거에 대한 조언 → 무조건 제거해야 하나? (필요성)
무조건 중복 이미지를 제거해야 하는지?
리얼 이미지 제거에 대한 조언
이미지 크기 차이에 대한 조언
만들어진 AI를 전이 학습 시에만 이용하지 못하는 건지, 전체적으로 이용 못하는 건지?
AI기술이 어떤 범위까지를 말하는 건지?
결과 제출 양식은 어떻게 되는지?


colab
드라이브
새로 만들기
내 드라이브
모델결과정리
Random Undersampling + 4배 Upsampling
Random Undersampling + 비율대로(2,3배) Upsampling
baseline model

목차




EDA

- 데이터 셋 살펴보기
- 문제점 도출
- 해결 방안

실사 이미지 제거

- 객체 탐지
- 군집화
- 필터링

데이터 불균형 해결

- Oversampling
- Undersampling
- 클래스 가중치

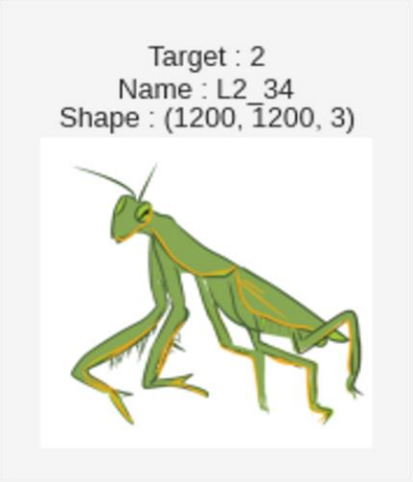
모델링 및 성능평가

- Resnet 기반 모델
- 커스텀 모델
- 성능평가

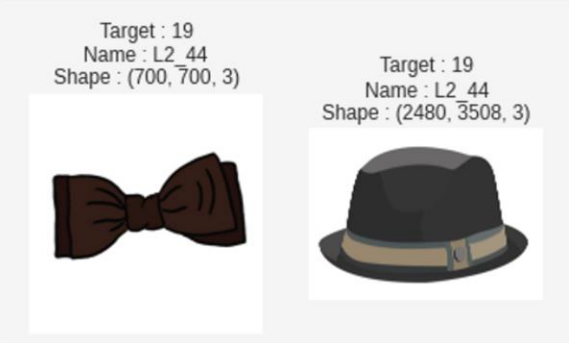
데이터 셋 살펴보기

L2_15: 전자기기

L2_30: 생활용품



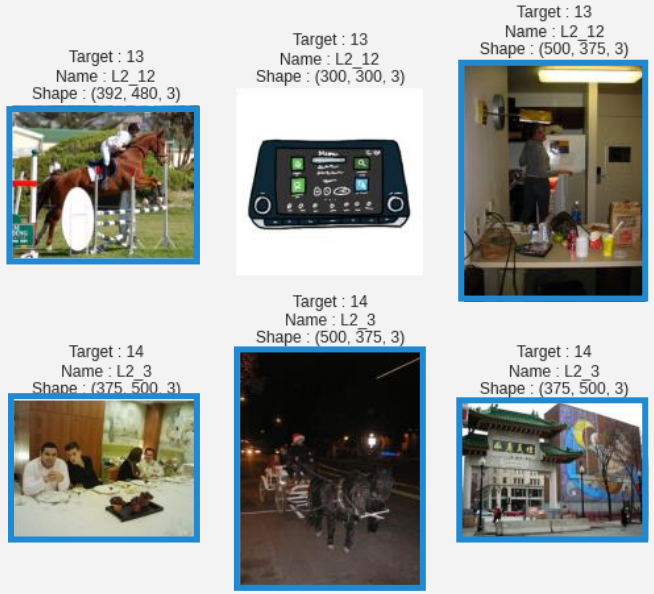
L2_34: 곤충



L2_52: 액세서리

이미지 데이터 **25503장, 클래스 20개**

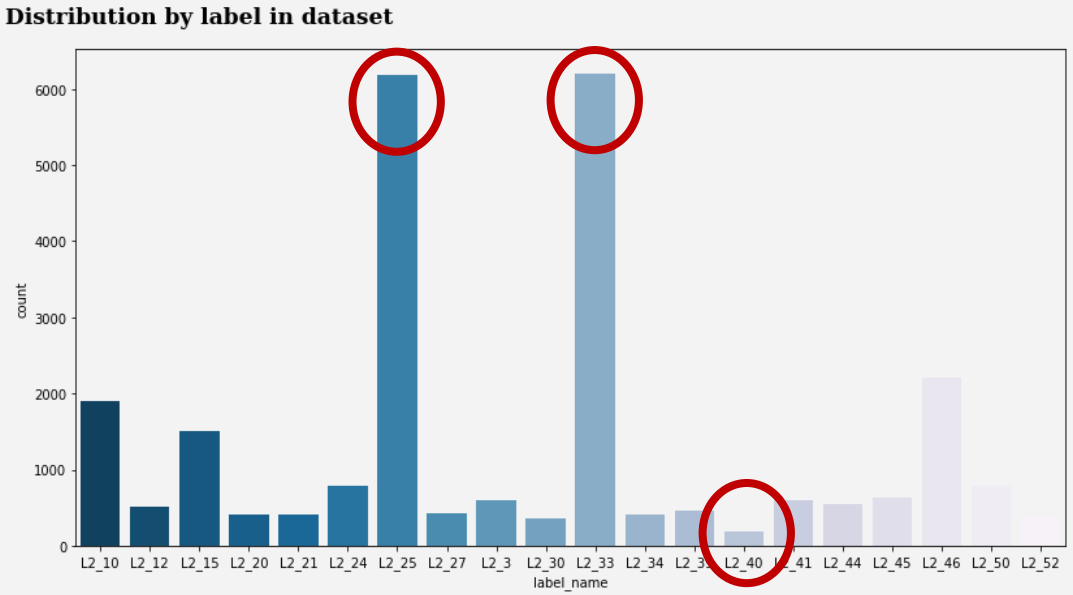
문제점 도출



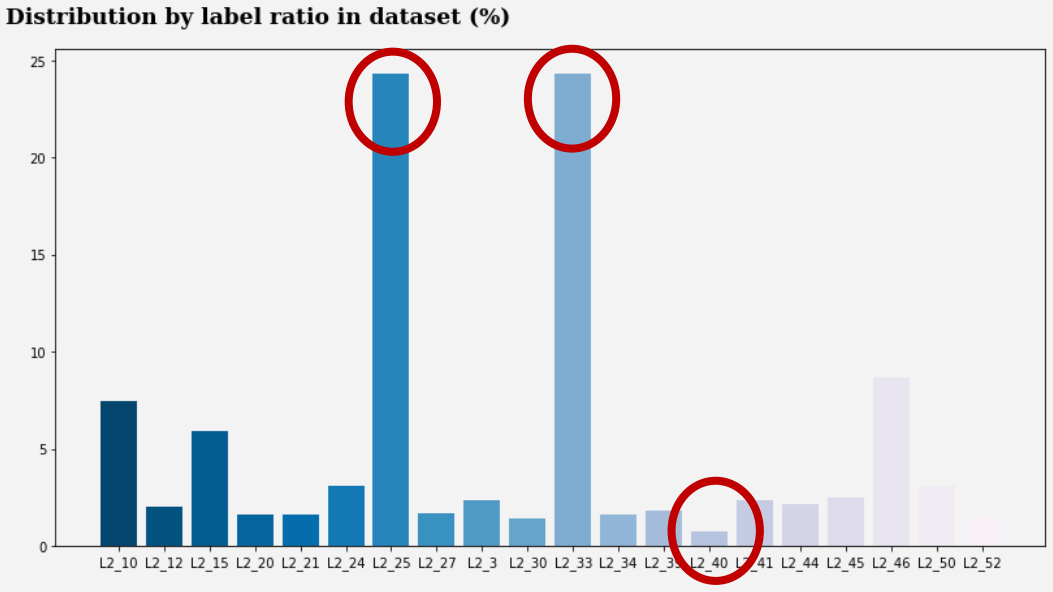
실사 이미지(Real Image)

중복 혹은 유사 이미지

문제점 도출



클래스별 데이터 개수 분포

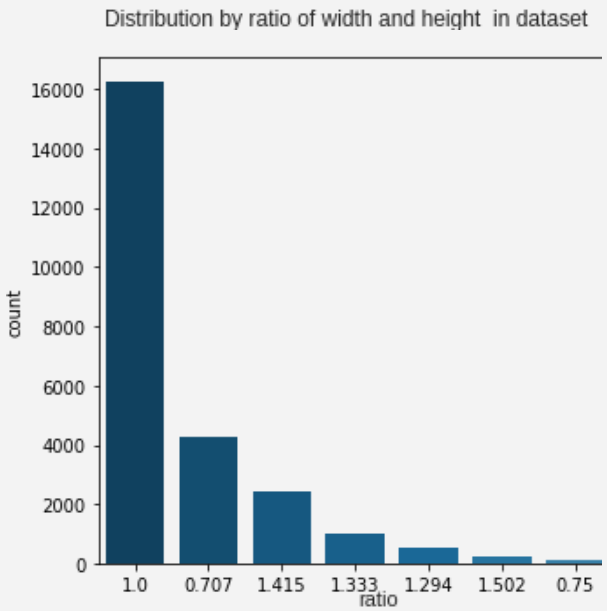


클래스별 데이터 비율 분포

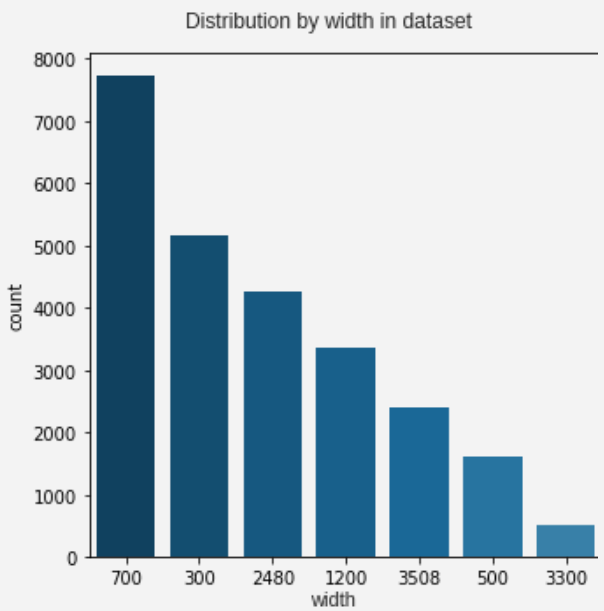
데이터 불균형

L2_33(6206개)가 클래스 L2_40(180개) 보다 **약 34배** 정도 많음

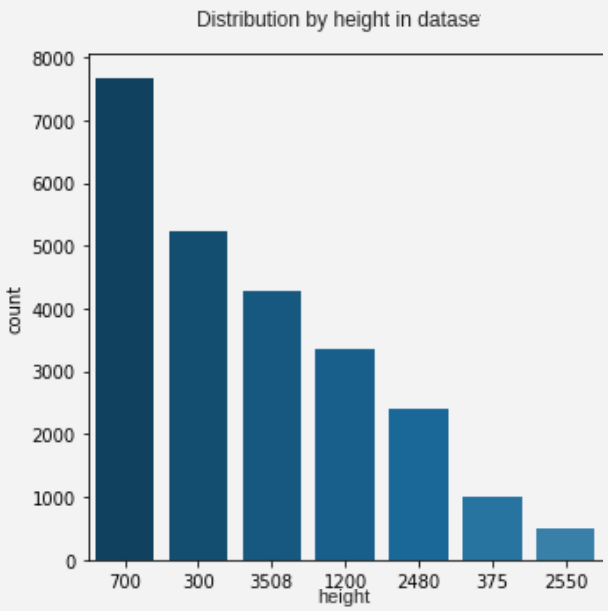
문제점 도출



이미지의 가로 세로 비율 분포



이미지의 가로 사이즈 분포



이미지의 세로 사이즈 분포

이미지별 사이즈 **불일치**

해결방안

문제점

실사 이미지 존재	중복 이미지 존재	클래스별 데이터 불균형	이미지별 사이즈 불일치

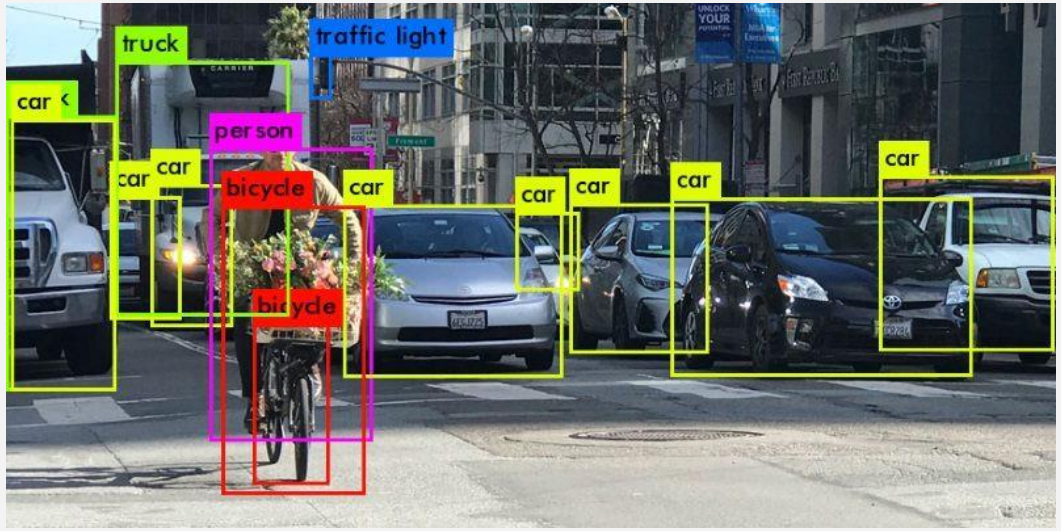
해결방안

- Object Detection - Clustering - Filtering	- 이미지 증강	- Oversampling - Undersampling - 클래스 가중치	- 파라미터 조절 (256,256)

실사 이미지 제거

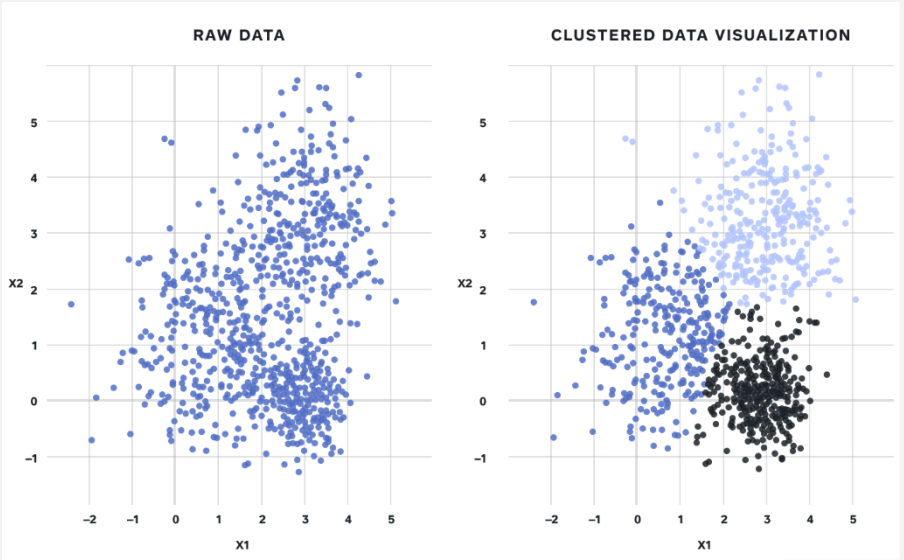
Solution 1 - Object Detection

Object Detection 모델 Yolo를 사용해
탐지되는 객체 수에 따라 이미지 제거



Solution 2 - Clustering

이미지의 Feature로 군집화하고
실사 이미지 군집은 제거



출처: https://developer.squareup.com/blog/so-you-have-some-clusters-now-what/
출처: https://heartbeat.comet.ml/introduction-to-basic-object-detection-algorithms-b77295a95a63?gi=c6809abf195f

객체 탐지

Solution 1 - Object Detection

Object Detection 모델 YOLO로
이미지를 Detect

객체가 1개일 경우: **일러스트**

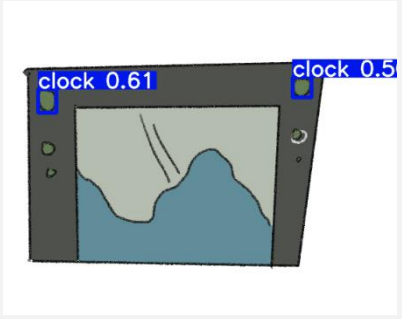
객체가 2개 이상일 경우: **실사 이미지**

객체 탐지

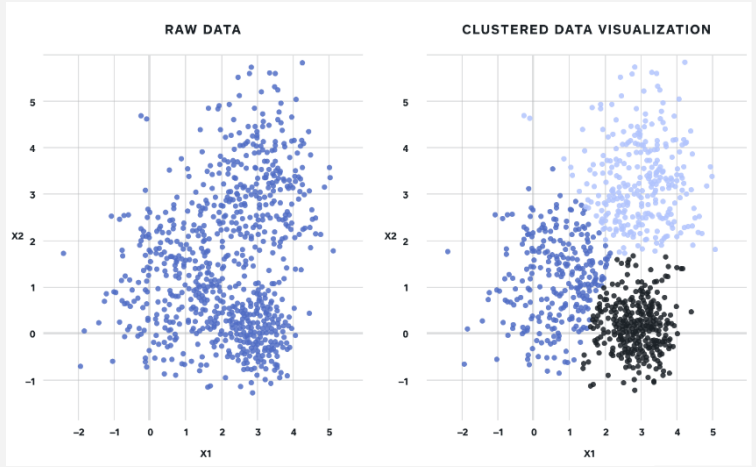
Solution 1 - Object Detection

객체가 1개지만 **실사 이미지**

객체가 2개 이상이지만 **일러스트**



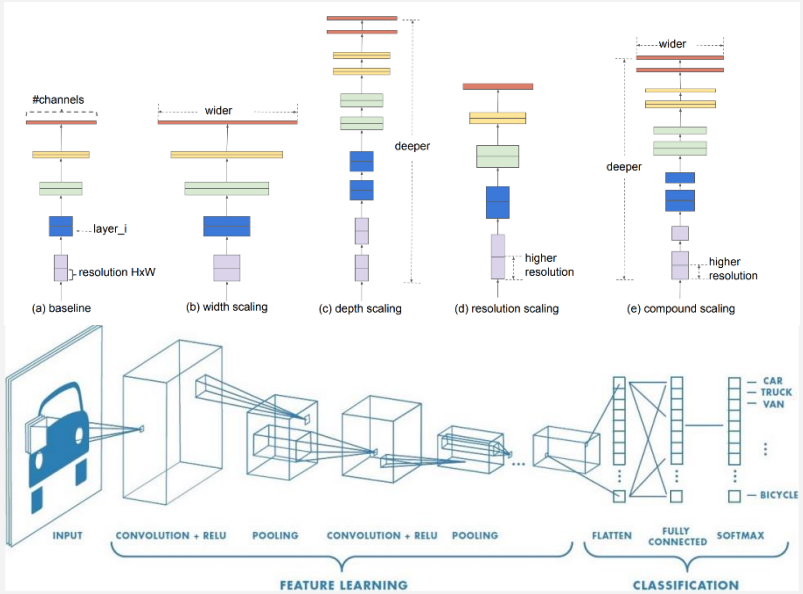
Solution 2
Clustering



일러스트인지? 실사 이미지인지?
라벨이 필요 없는 비지도 학습

군집화

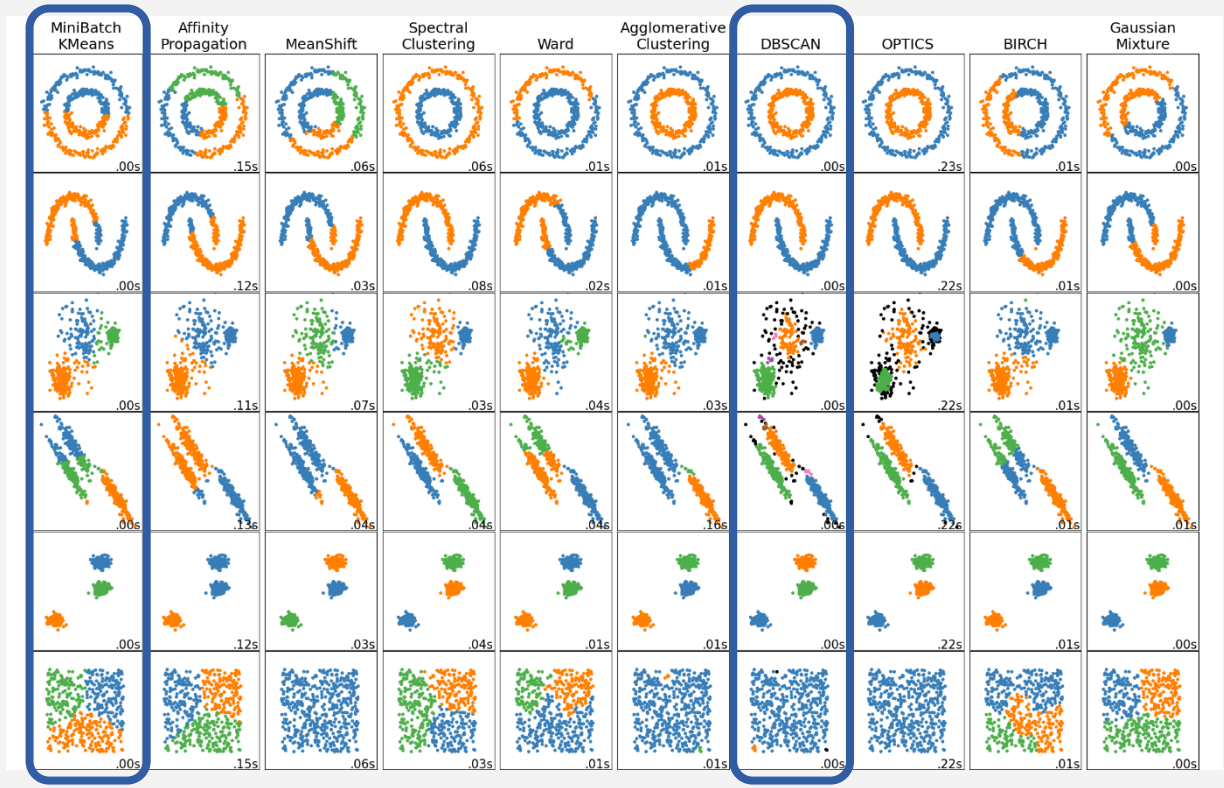
Solution 2 - Clustering



Efficientnet b4

CNN based
Feature Extractor

KMeans DBSCAN



Clustering model

과학기술정보통신부 NIA 한국지능정보사회진흥원

출처 https://theaisummer.com/static/8d19d048cd68d6dce362e025cf3b635a/1ac66/skip-connection.png

군집화 모델 비교

Solution 2 - Clustering

Best Clustering Model

KMeans < DBSCAN

WHY?

KMeans

구역을 기준으로
Clustering

DBSCAN

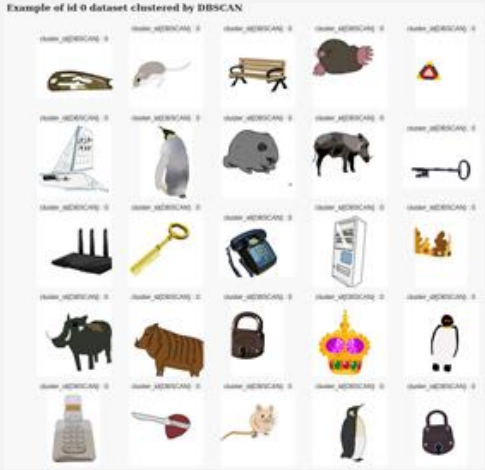
각 원소들의 군집 밀도에 따라
Clustering

DBSCAN

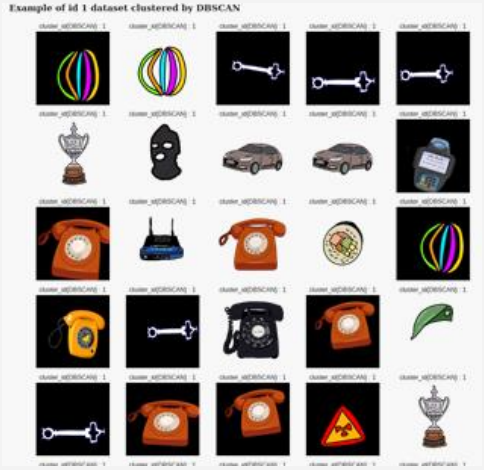
Solution 2 - Clustering

DBSCAN 결과

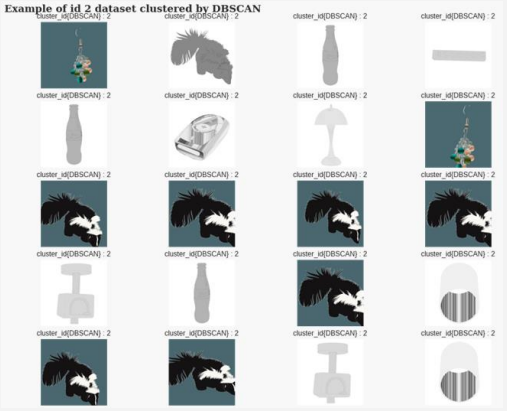
Cluster -1



Cluster 0



Cluster 1



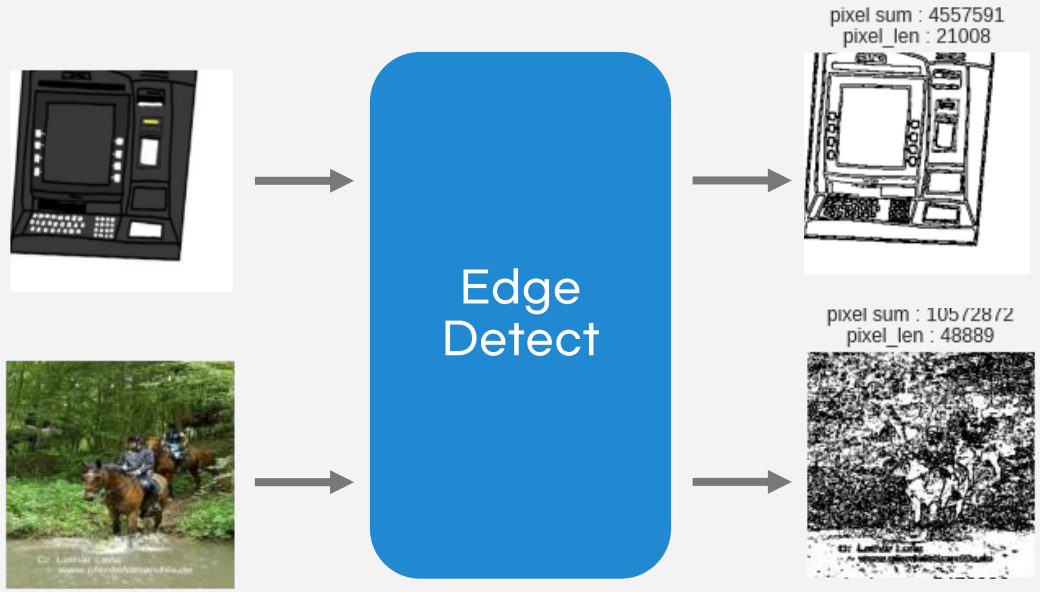
Cluster 2

실사 이미지들은
Cluster 되지 못함

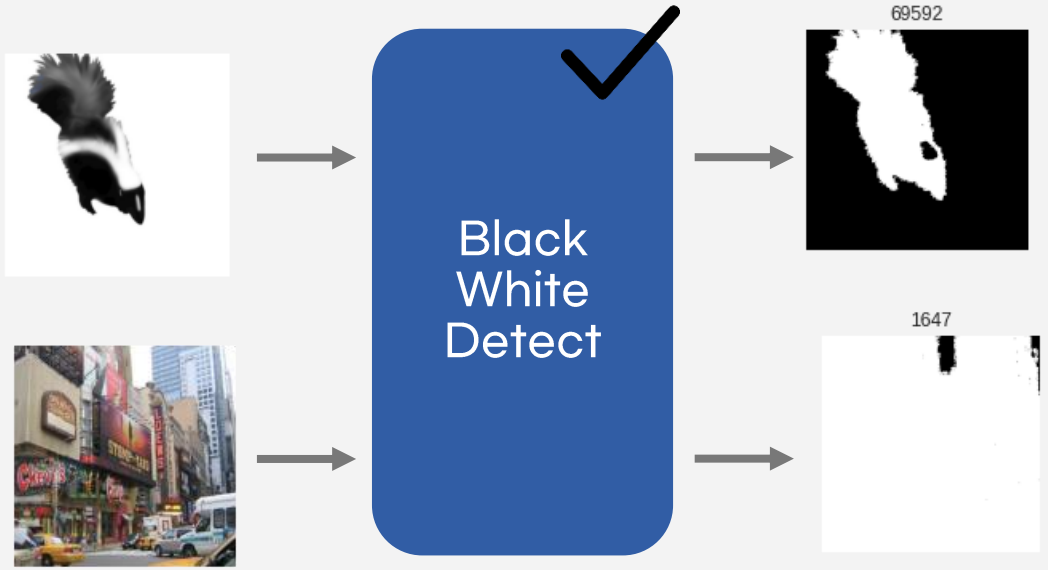
필터링

Solution 2 - Clustering + Filtering

Edge Detection Filtering

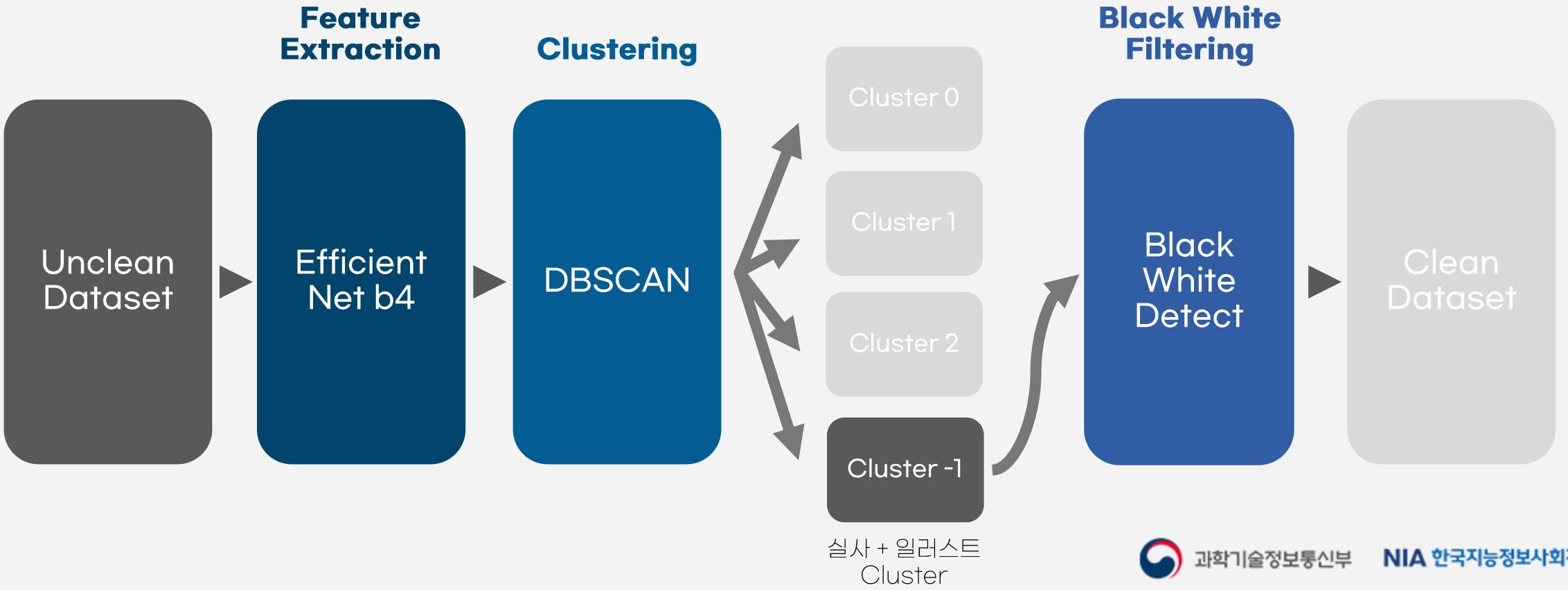


Black White Filtering



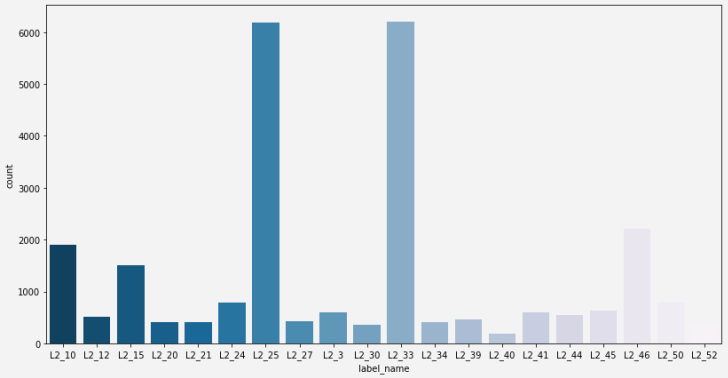
최종 실사 이미지 제거

Solution 2 - Clustering + Filtering



데이터 불균형 문제 해결

클래스별 데이터 분포



25,300 → 23,000

원본 데이터셋

실사이미지 제거후

Imbalanced Data 해결 아이디어

Oversampling
이미지 증강

Undersampling
랜덤 샘플링

Class Weight
가중치 적용

Oversampling

이미지 증강 회전,뒤집기,대비/밝기 변화,이미지 크기 변화,크롭

```
def albu_transforms():
    return A.Compose([
                A.Rotate(),
                A.OneOf([
                        A.HorizontalFlip(p=1.0),
                        A.VerticalFlip(p=1.0),
                        ], p=0.5),
                A.OneOf([
                        A.RandomContrast(limit=0.5, p=1.0),
                        A.RandomGamma(gamma_limit=(200, 250), p=1.0),
                        A.RandomBrightnessContrast(brightness_limit=0.4, contrast_limit=0.4, p=1.0,),
                        ], p=1.0),
                A.OneOf([
                        A.Resize(256, 256),
                        A.Compose([
                            A.Resize(320, 320),
                            A.CenterCrop(256, 256),
                        ])
                        ], p=1.0)
                ])
```

클래스별 증강 배수 설정

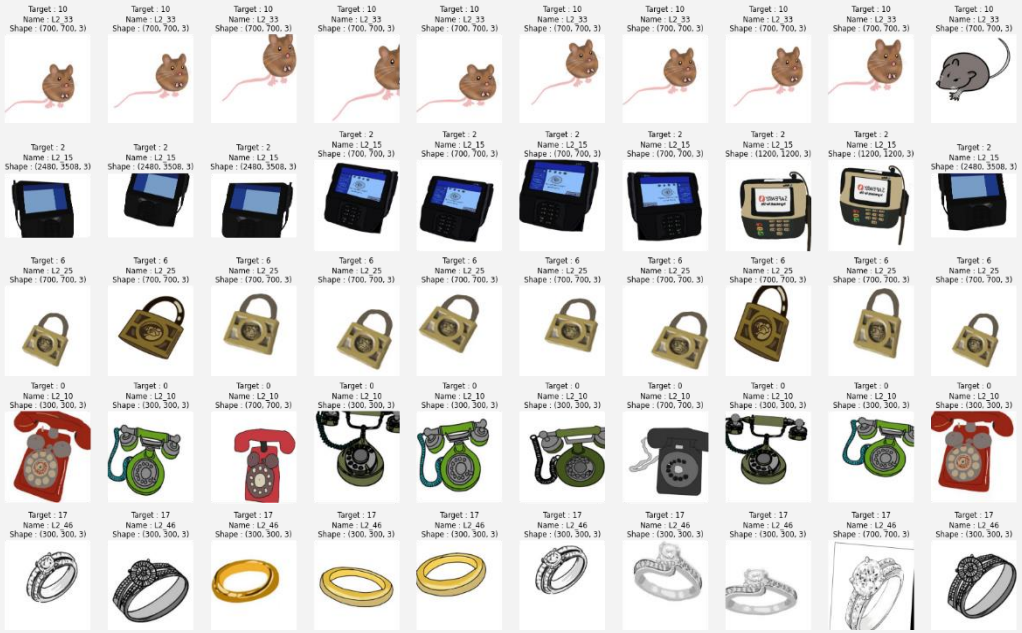
데이터가 매우 적은 클래스
L2_12 X 3

데이터가 적은 클래스
L2_3, L2_40, L2_41 X 2

Undersampling

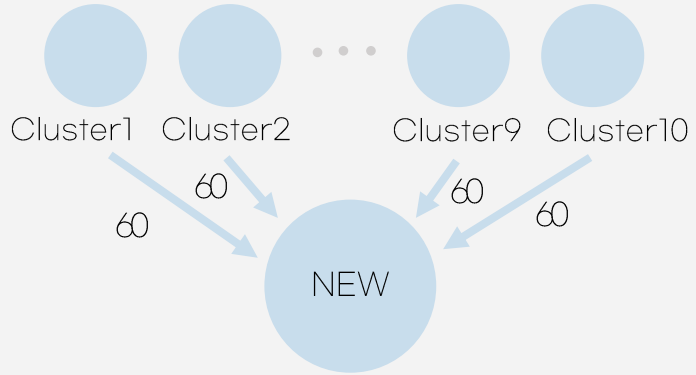
데이터가 평균 이상인 클래스

L2_15, L2_46, L2_25, L2_10, L2_33



Solution 1 - Clustering

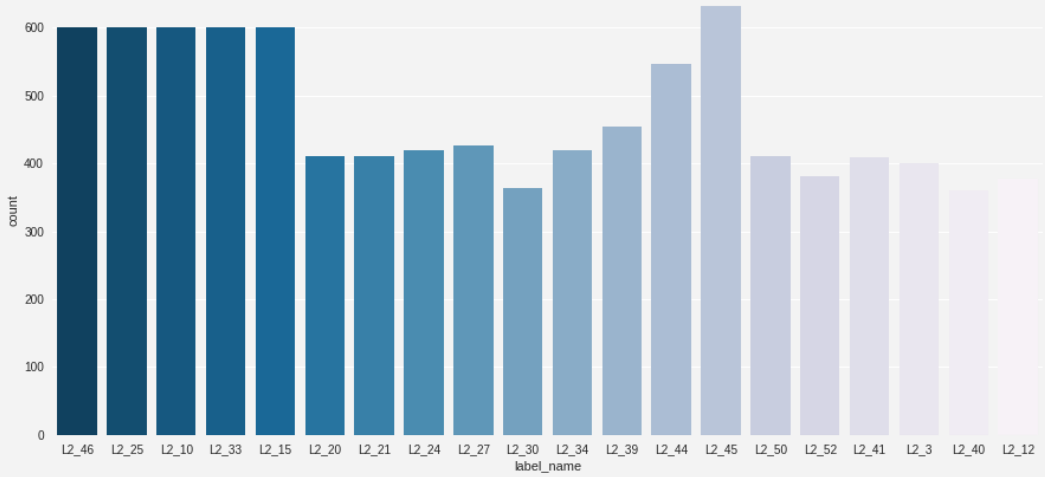
K=10 군집에서 군집별 60개씩 데이터 샘플링



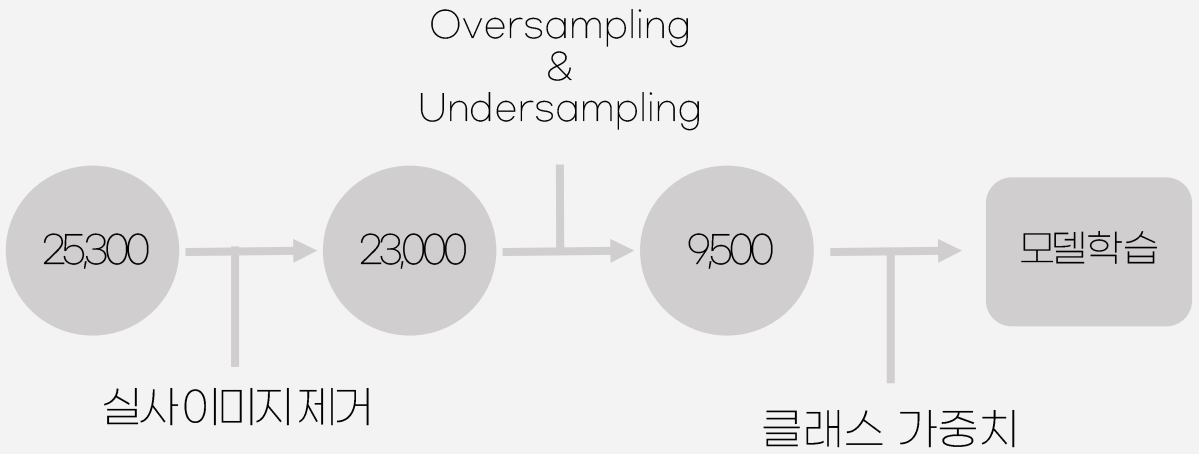
Solution 2 - Random sampling

클래스별 랜덤 600개씩 데이터 샘플링 ☑

클래스 가중치



불균형 문제 해결 후 데이터 분포 변화

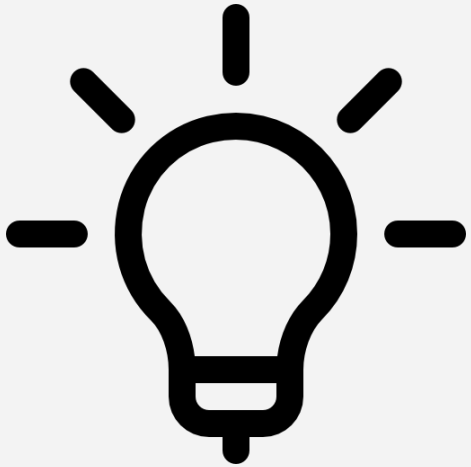


모델 학습 시 클래스 가중치 적용

```
class_weights = compute_class_weight(class_weight = "balanced",
                                     classes = np.unique(df['label_name']),
                                     y = df['label_name'])
class_weights = dict(enumerate(class_weights))

history = model5.fit_generator(train_dataset, validation_data = validation_dataset,
                               epochs=n_epoch, class_weight = class_weights, callbacks = [e
```

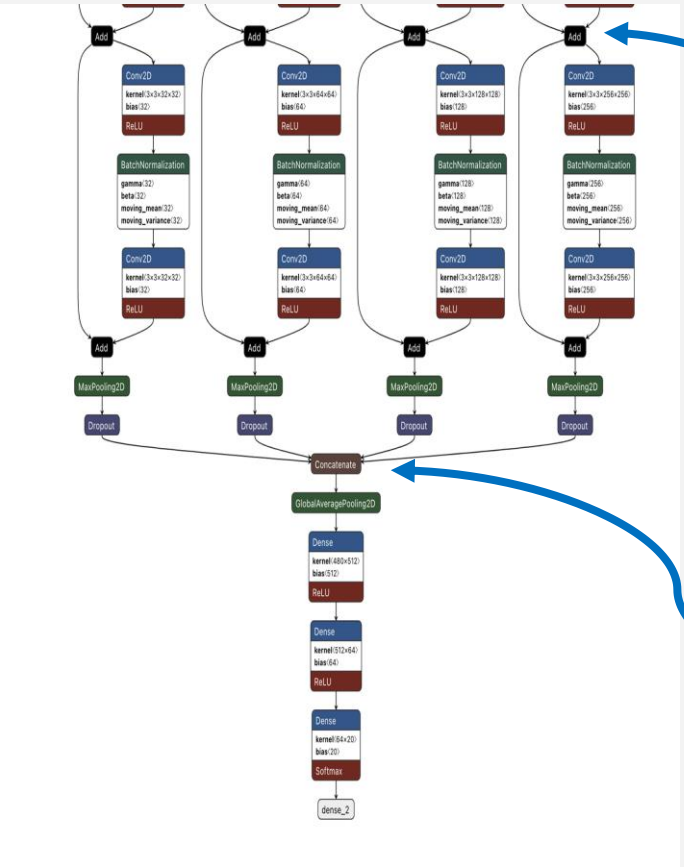
모델 설계



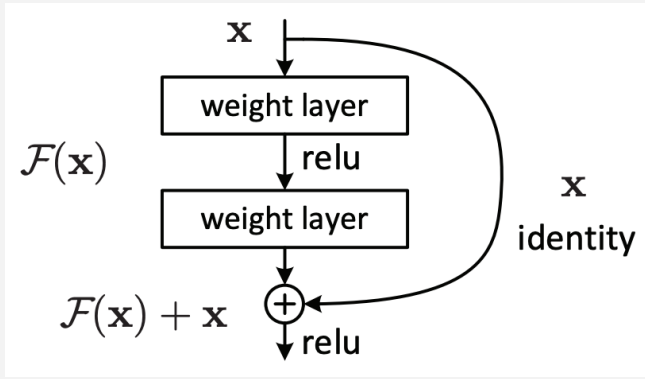
① ResNet, InceptionNet
아이디어 모델

② CNN
커스텀 모델

ResNet, InceptionNet 아이디어

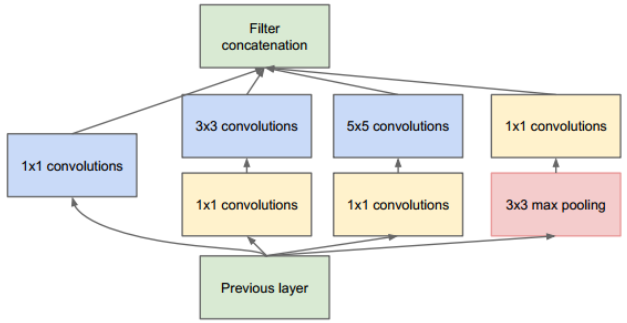


Residual Block



잔여효과로 인해
학습이 더 쉬워지고
기울기 소실 문제 완화

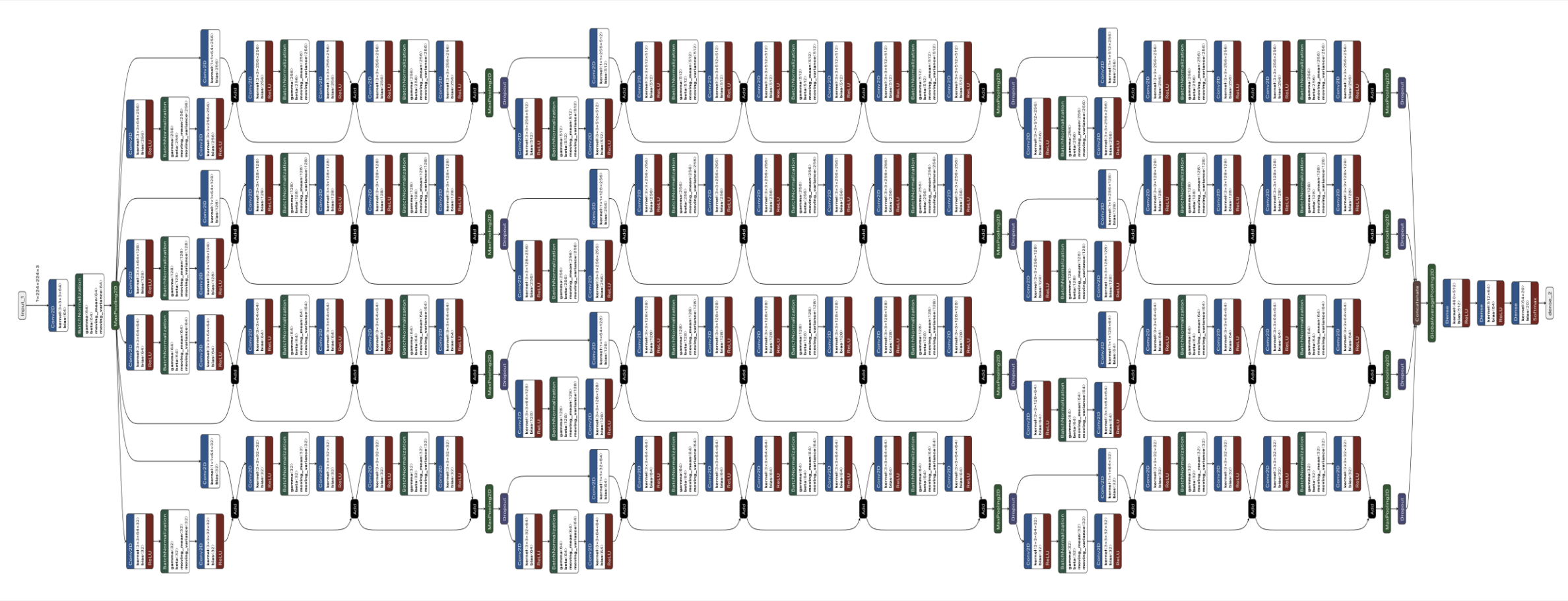
Inception Block



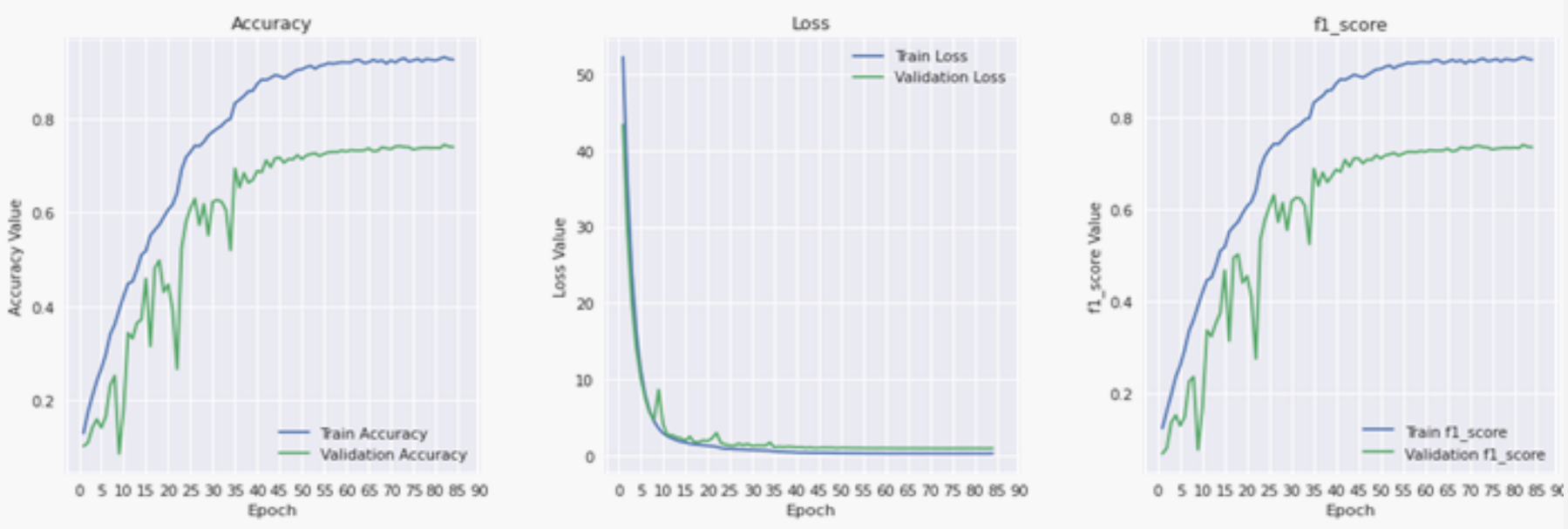
다양한 채널에서
추출된 Feature를
결합할 수 있음

출처: https://arxiv.org/abs/1512.03385 https://arxiv.org/abs/1409.4842

ResNet, InceptionNet 아이디어 모델

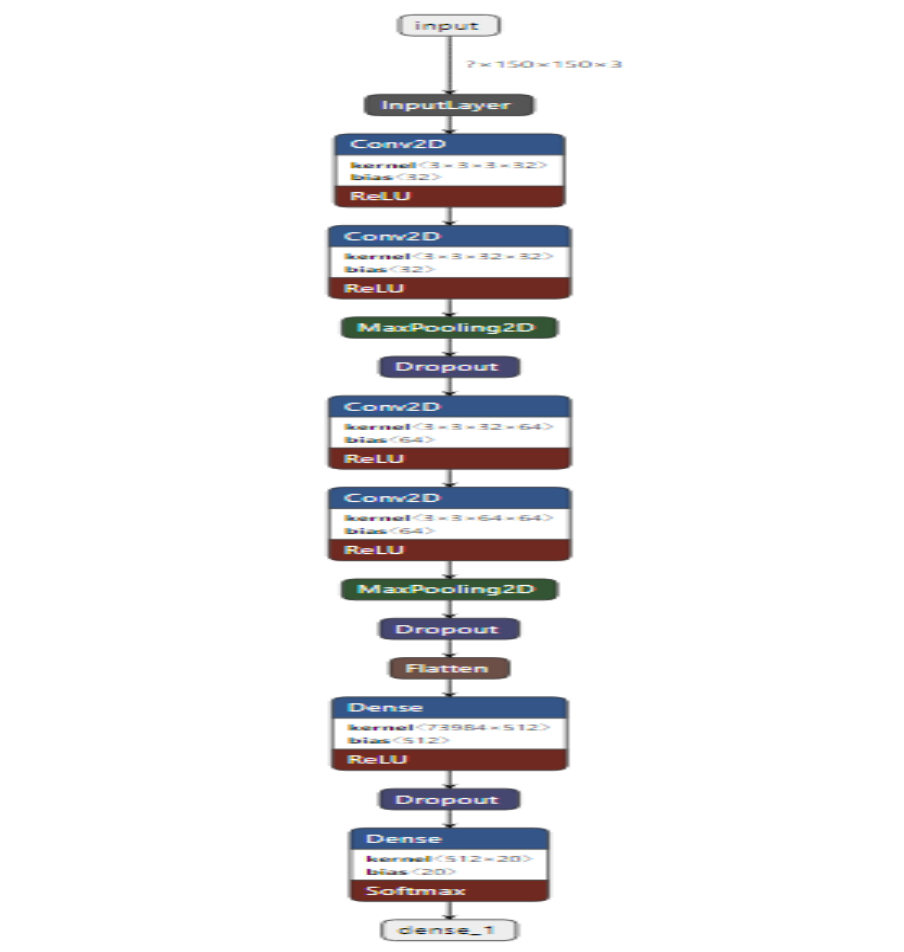


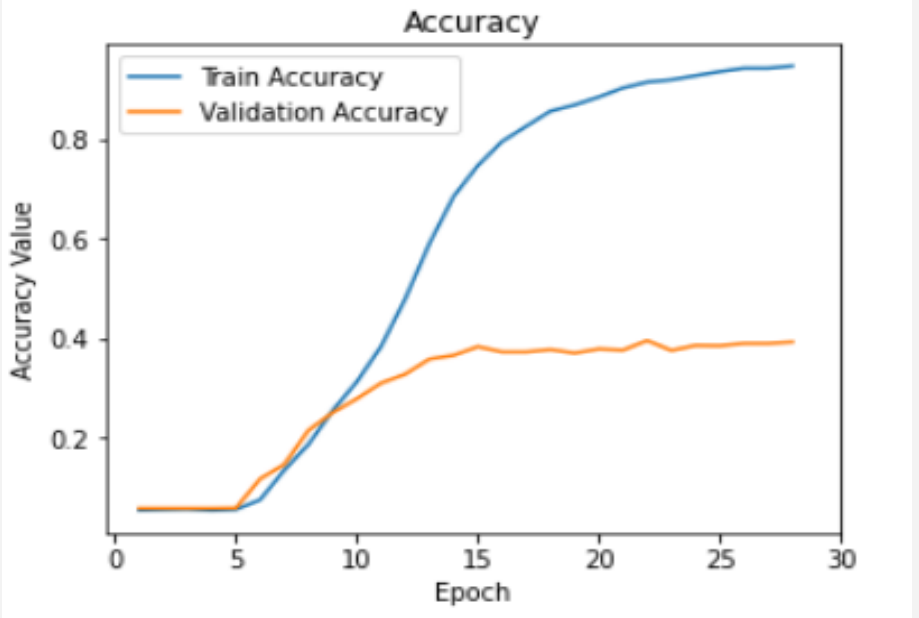
ResNet, InceptionNet 아이디어 모델 성능평가



F1 score = 0.7337

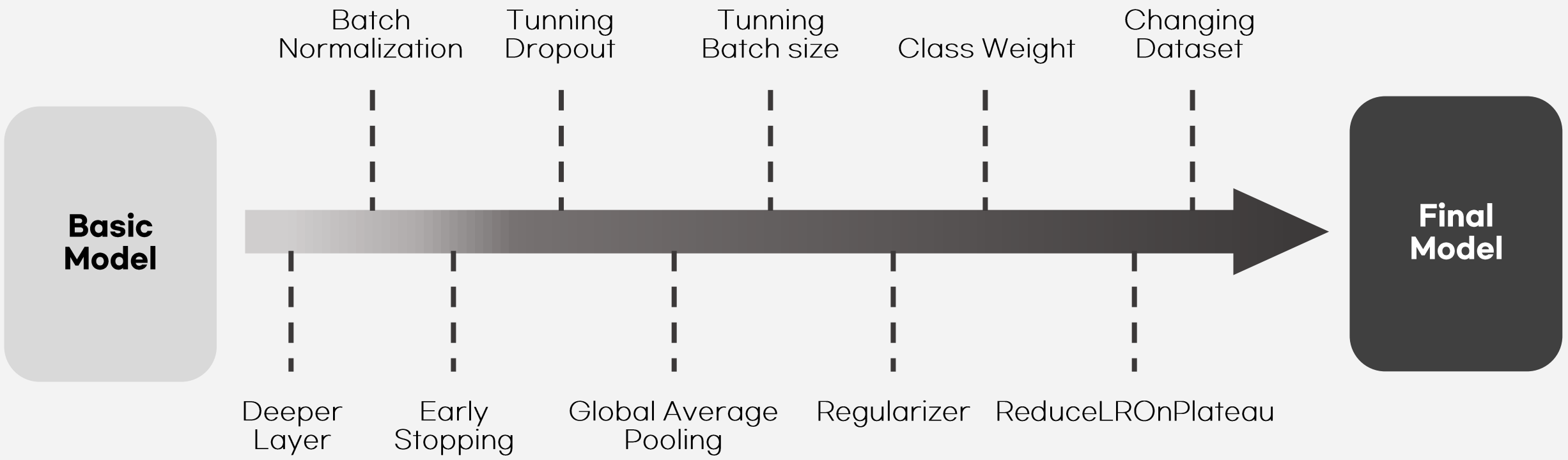
CNN 커스텀 모델 - Basic



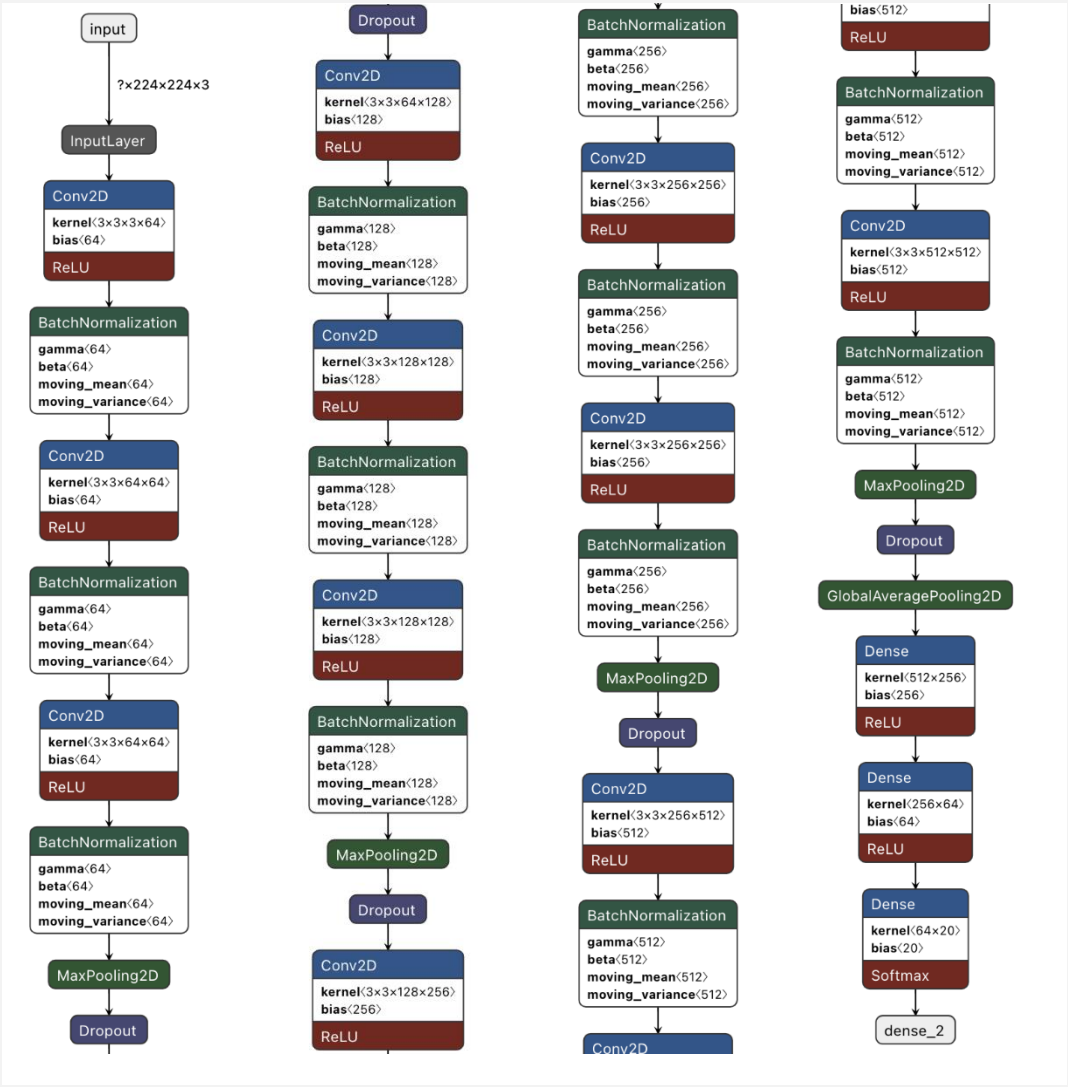


CNN 커스텀 모델 실험 전
기준 모델 생성

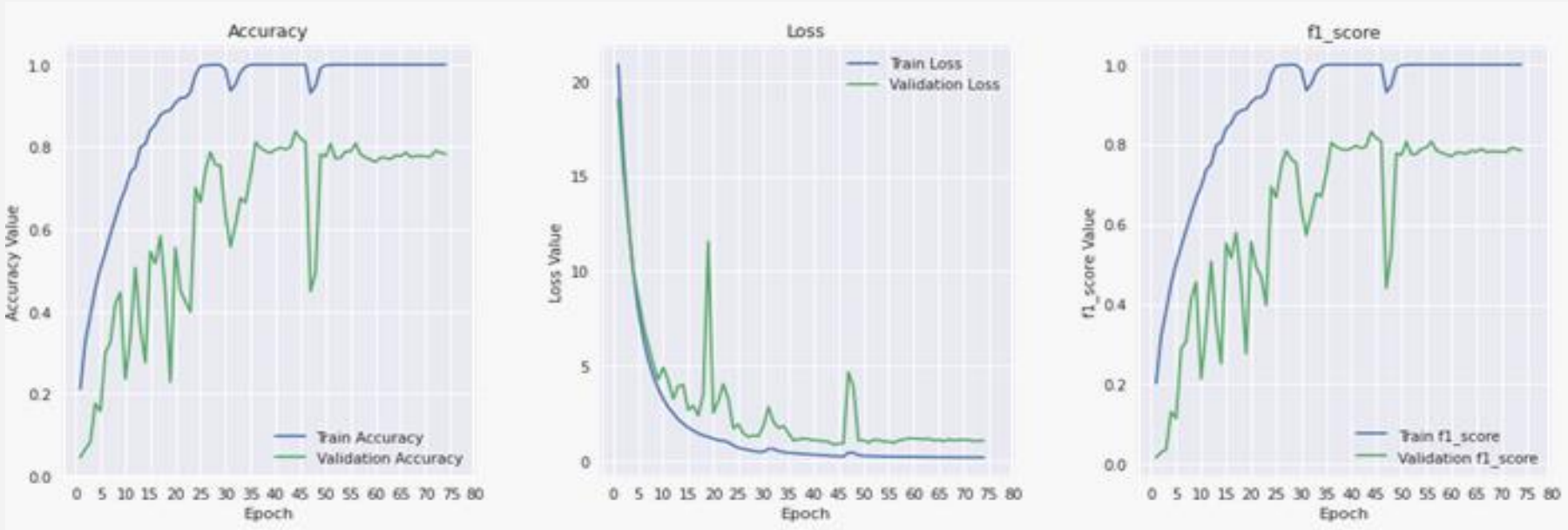
CNN 커스텀 모델


Basic Model
Batch Normalization
Tunning Dropout
Tunning Batch size
Class Weight
Changing Dataset
Deeper Layer
Early Stopping
Global Average Pooling
Regularizer
ReduceLROnPlateau
Final Model

CNN 커스텀 모델 - Final



CNN 커스텀 모델 성능평가



F1 score = 0.8317

감사합니다

2022 DATA CREATOR CAMP

과학기술정보통신부 NIA 한국지능정보사회진흥원