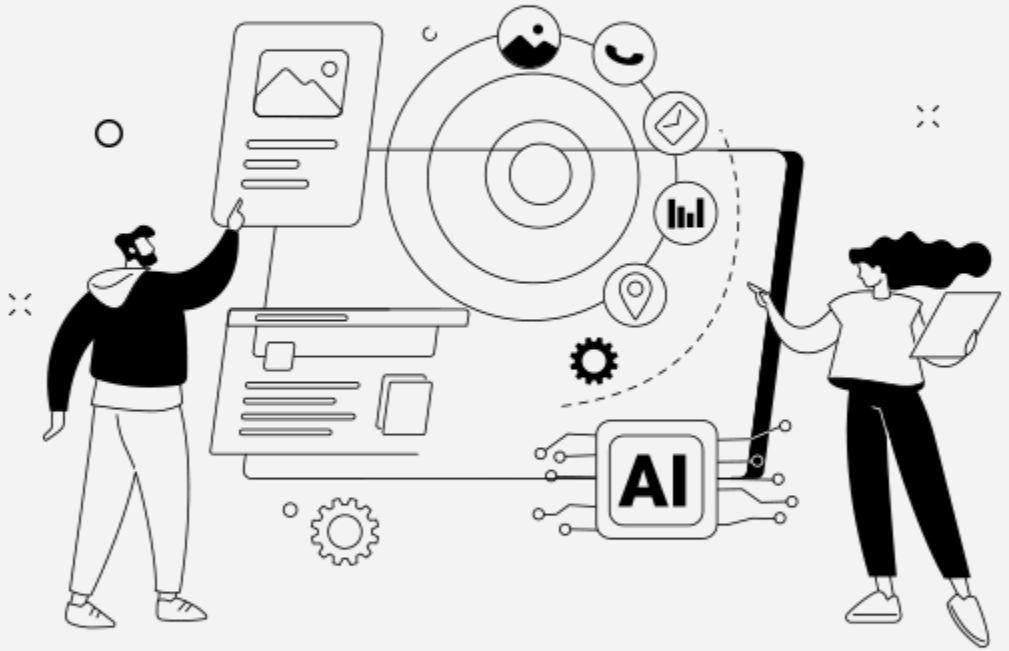
2022 데이터 크리에이터 캠프

Data Creator Camp



4회차 비포

윤병효(팀장) 김혜인 정다인 정솔잎

과학기술정보통신부 NIA 한국지능정보사회진흥원

Contents

A

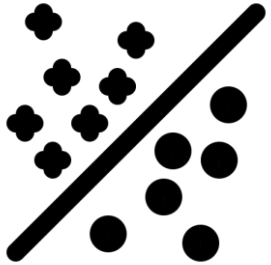
Mission 1

탐색적 자료 분석(EDA)

학습 영상에 대한 분포, 특성 확인

데이터 분포 관련 문제 해결

B

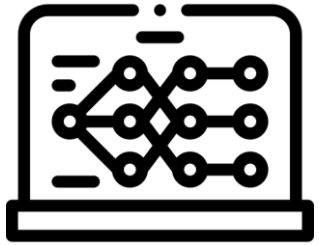


Mission 2

AI 기반 영상 제거

학습데이터에 포함된 사진 영상 제거

C



Mission 3

신경망 네트워크 모델링

데이터 전처리, 영상처리, 파라미터

설계, 성능 최적화, F1 score

Mission 1.

EDA - 클래스 (라벨) 별 분포 및 특성 확인

	name	label	width	height	pixel
0	ycrqupsfbtdmppsedlxl.jpg	L2_10	700	700	175.475408
1	oqfadnuqsmolzmxwfycm.jpg	L2_10	300	300	187.451700
2	edkxyljaevluzpccthfb.jpg	L2_10	700	700	204.499047
3	qfcrkaiksvpsezumhlvi.jpg	L2_10	300	300	191.282344
4	rtfhhjiiawrrdtcxjxyz.jpg	L2_10	700	700	195.024749
...	...	...	...	...	...
25498	fbkloojqdqzzvyzjhrfv.jpg	L2_52	3508	2480	228.673894
25499	bzzgdseklzecfcoujbaf.png	L2_52	700	700	240.611075
25500	cpexnivtruoevtlazloq.jpg	L2_52	3508	2480	204.656396
25501	dtxfpcggpnrgqolnmdoy.png	L2_52	700	700	234.191987
25502	yifaakzphsggoouohtan.jpg	L2_52	300	300	229.895689

25503 rows × 5 columns

이미지에서 특성값 width(가로),

height(세로), pixel(픽셀) 추출

+

이미지 채널을 BGR에서 RGB로 변경

+

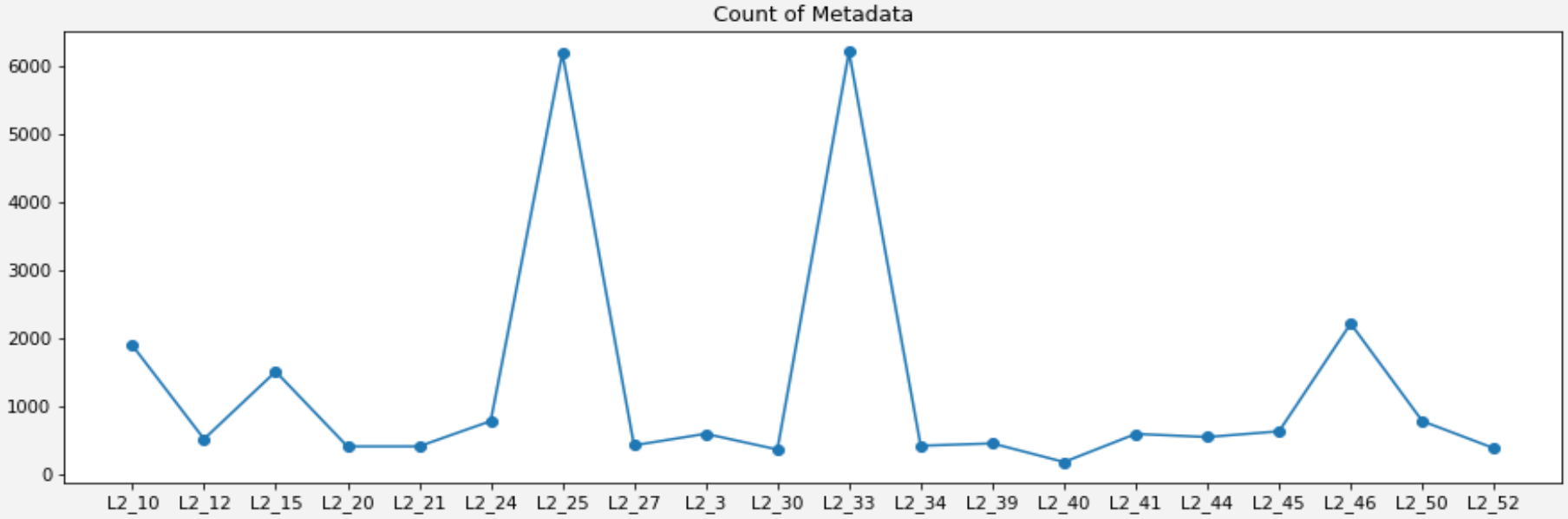
각 이미지마다 픽셀 평균값 계산

↓

Metadata 데이터 프레임 생성

Mission 1.

EDA - 클래스 (라벨) 별 분포 및 특성 확인



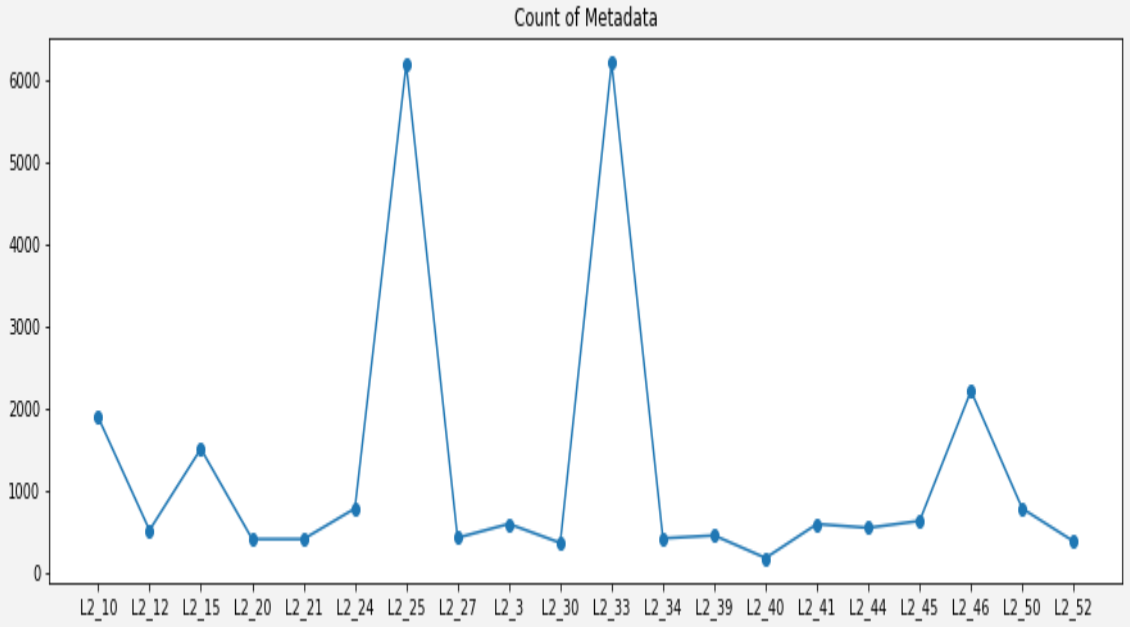
클래스 L2_25, L2_33의 데이터 개수가 각각 6189, 6206개로 다른 클래스에 비해

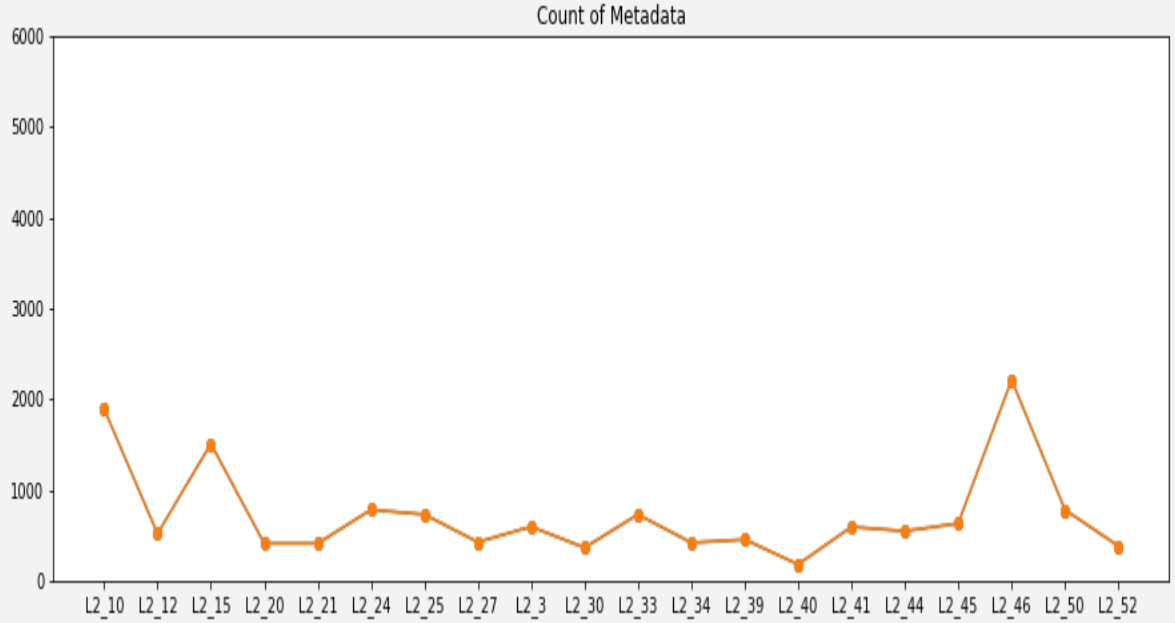
학습 데이터의 양이 과도하게 많고, L2_40 의 데이터 수는 180개로 매우 적음

"학습 데이터의 **분포 불균형** 문제가 있다고 판단"

Mission 1.

EDA - 데이터 불균형 분포 해결

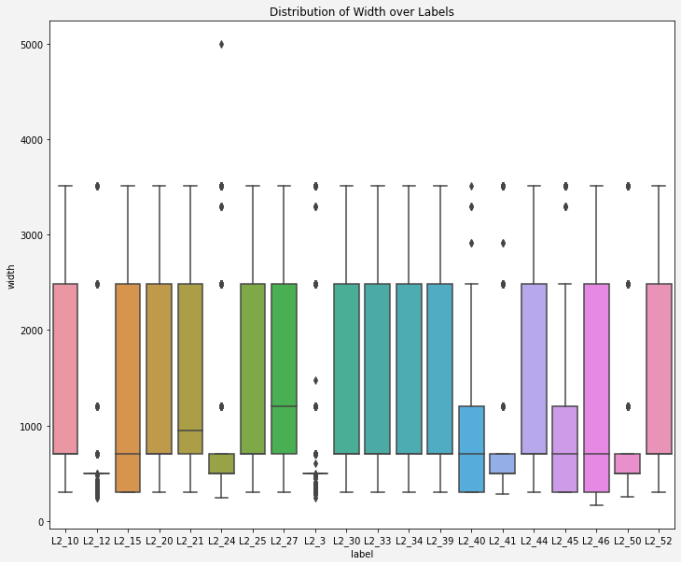




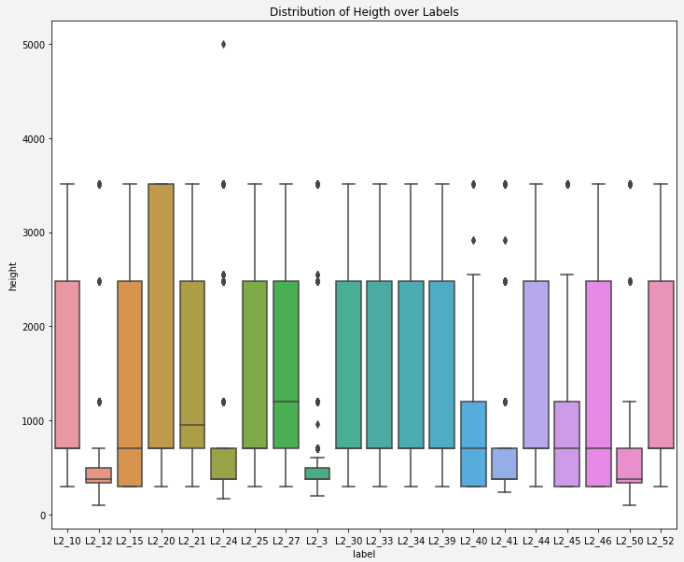
클래스 L2_25, L2_33이 다른 클래스에 비해 데이터 개수가 많음. 나머지 클래스 정보 손실을 막기 위해 두 클래스만 undersamlping 진행. 두 클래스를 제외한 나머지 클래스의 평균 값을 구해 undersamlping 개수를 지정하고 두 클래스에 적용

Mission 1.

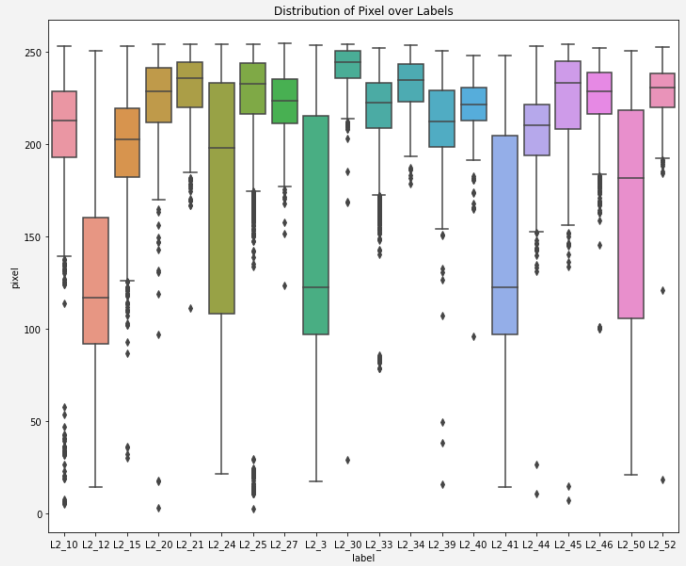
EDA - 클래스 (라벨) 별 분포 및 특성 확인



클래스별 width 분포



클래스별 height 분포



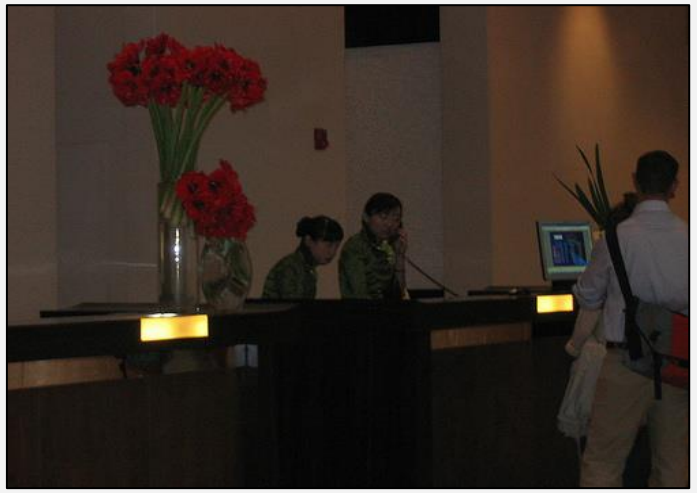
클래스별 픽셀 평균값 분포

클래스 L2_12, L2_24, L2_3, L2_41, L2_50은 다른 클래스들에 비해 상대적으로 평균값이 낮음.

Mission 1/2.

EDA - 일러스트 이미지 vs 실제 사진 이미지

일러스트 이미지



실제 사진 이미지

```
np.mean(np.array(illust))
220.18781276437267
```

```
np.mean(np.array(real))
41.11399822222222
```

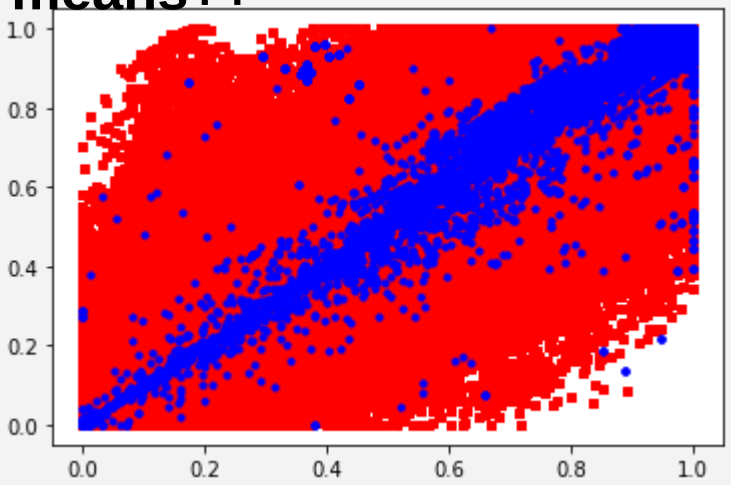
일러스트 이미지의 경우, 흰 배경(255)값이 대부분이라 픽셀 평균값이 높음.
반면, 실제 사진(오염된 이미지는)는 픽셀 평균값이 낮게 나옴

클래스 L2_12, L2_24, L2_3, L2_41, L2_50은 다른 클래스들에 비해 **pixel 평균값이 낮은 것으로** 보아 **해당 클래스에 오염된 데이터들이 들어가 있을 가능성**이 있다고 판단

Mission 2.

AI 기반 영상 제거 – k-means++

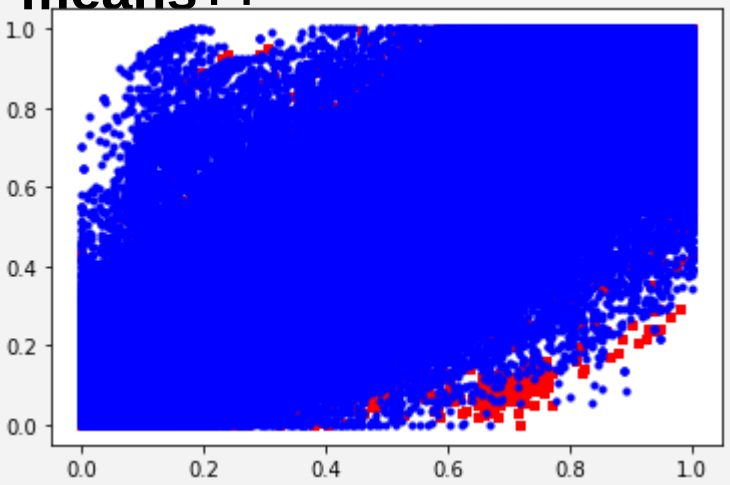
임의 차원 축소 + k-means++




```
print(data[y_km == 0].shape)   #실사
print(data[y_km == 1].shape)   #실사 아닌 것

(1905, 224, 224, 3)
(1364, 224, 224, 3)
```

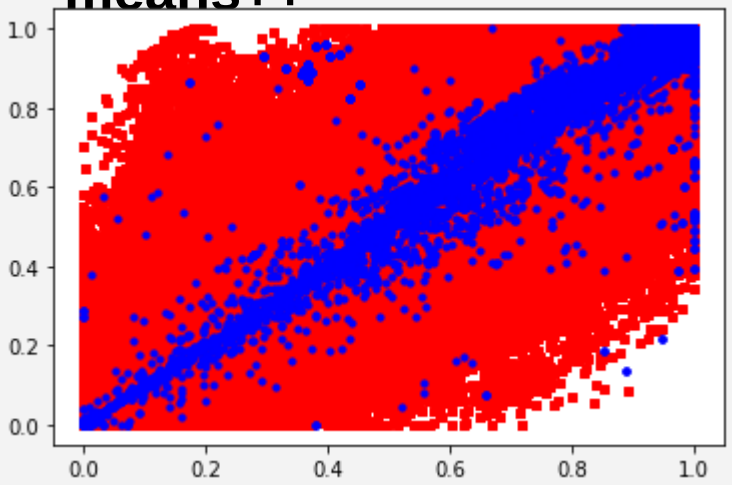
PCA 차원 축소 + k-means++




```
print(data[y == 0].shape)  #실사
print(data[y == 1].shape)  #실사 아닌 것

(327, 224, 224, 3)
(2942, 224, 224, 3)
```

AE 차원 축소+ k-means++



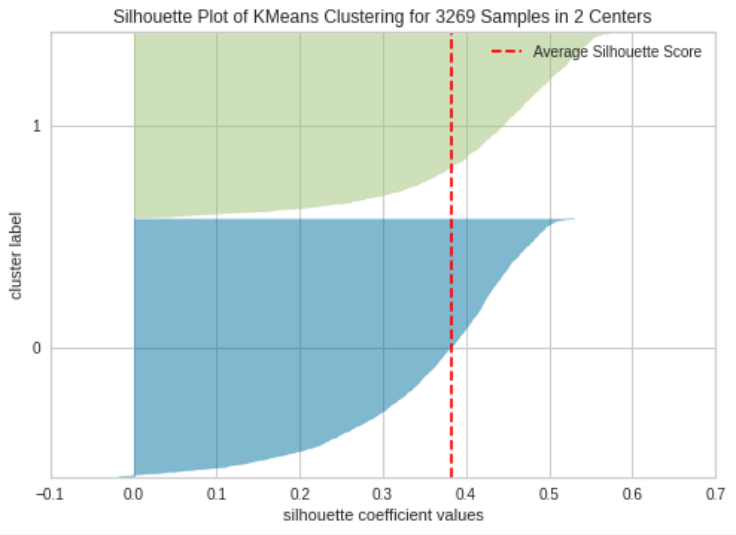

```
print(data[y_km == 0].shape)   #실사
print(data[y_km == 1].shape)   #실사 아닌 것

(1905, 224, 224, 3)
(1364, 224, 224, 3)
```

Mission 2.

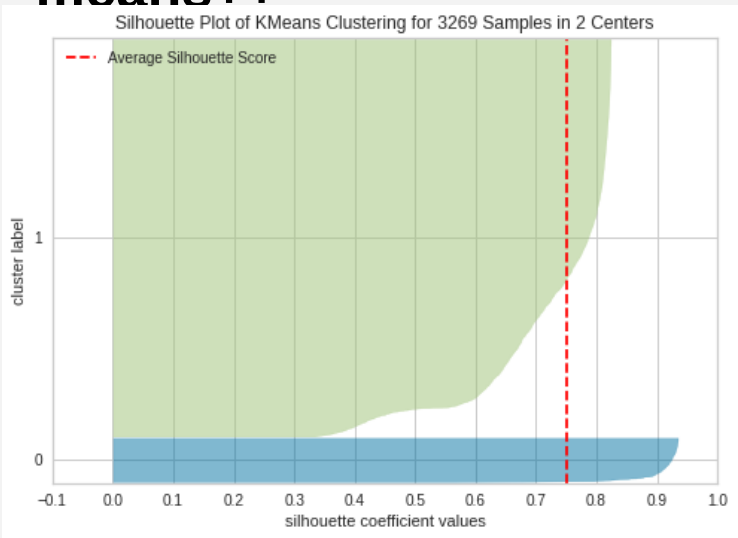
AI 기반 영상 제거 - 평가(실루엣 계수, DBI)

임의 차원 축소 + k-means++



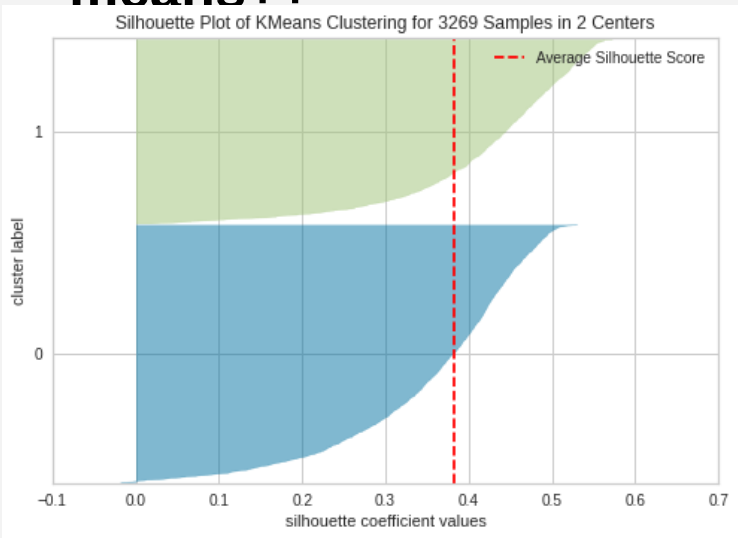
Silhouette Coefficient: 0.3829
Davies Bouldin Index: 1.0540

PCA 차원 축소 + k-means++



Silhouette Coefficient: 0.0048
Davies Bouldin Index: 14.7660

AE 차원 축소+ k-means++



Silhouette Coefficient: 0.3829
Davies Bouldin Index: 1.0540

Mission 2.

AI 기반 영상 제거 - 모델 선정 및 실사 이미지 제거

Cluster = 0

Cluster = 1

실루엣 계수가 상대적으로 높고,

DBI가 상대적으로 낮은 1,3번 우수하다 판단

실사 이미지가 잘 분류된 것을 확인할 수 있음

↓

클러스터 0으로 분류된 이미지 제거 !

Mission 3.

모델링 – Train/Test data set split 설정

```
#train과 test데이터 8:2의 비율로 분리
X_train, X_test, y_train, y_test = train_test_split(df, df['label'].values, test_size=0.2)
print("Number of posters for training: ", len(X_train))
print("Number of posters for validation: ", len(X_test))
```

```
Number of posters for training:  12745
Number of posters for validation:  3187
```

Train과 Test 데이터셋을 8:2의 비율로 분리한 결과,
Train Dataset은 12745개, Test Dataset은 3187개로 분리

Mission 3.

모델링 – Data Augmentation, validation data set split 설정

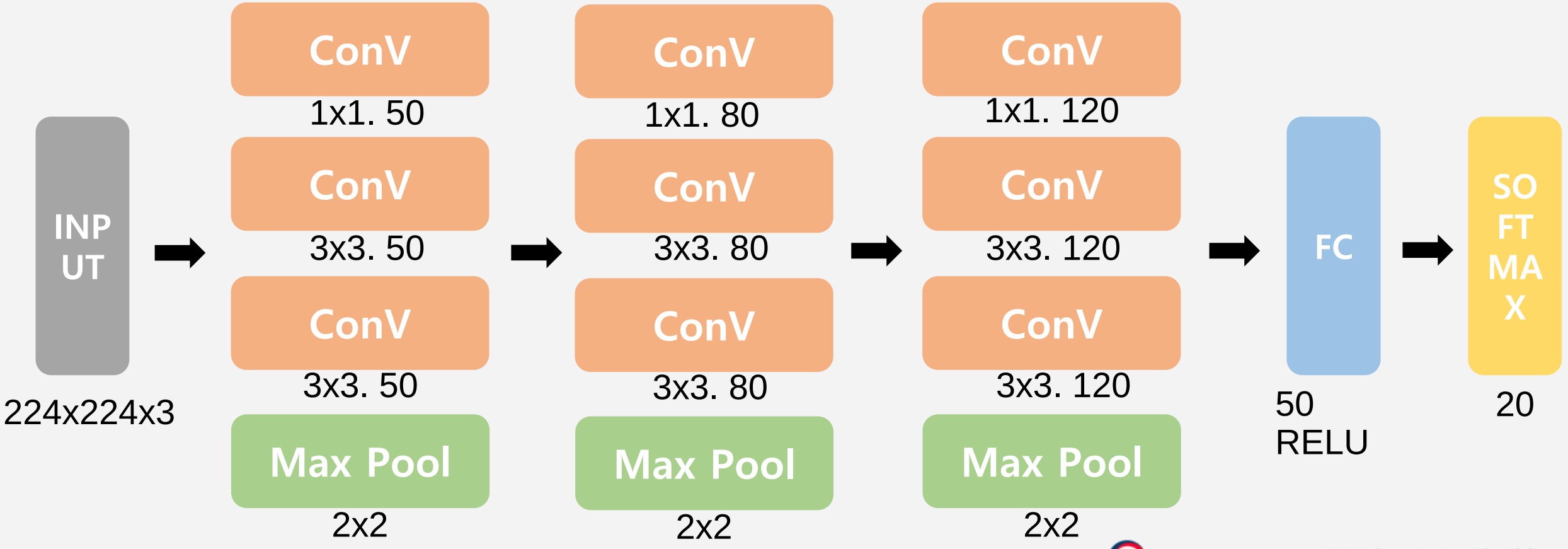
```
DATAGEN_TRAIN = ImageDataGenerator(
    rescale=1./255,
    rotation_range=10,
    width_shift_range=0.1,
    height_shift_range=0.1,
    shear_range=0.2,
    zoom_range=0.2,
    horizontal_flip=True,
    vertical_flip=True,
    data_format="channels_last",
    validation_split=0.3)
```

rescale : 1/255로 스케일링하여 0~1범위로 변환
rotation_range : [-10, 10] 각도 회전
width_shirt_range : width의 0.1픽셀 내외로 좌우 이동
height_shirt_range : height의 0.1픽셀 내외로 상하 이동
shear_range : [-0.1,0.1] 굴절
zoom_range : [0.8, 1.2] 확대 축소
horizontal_flip : 좌우 반전
vertical_flip : 상하 반전

train : validation = 7 : 3 비율로 나눔

Mission 3.

모델링 – CNN 구조



Mission 3.

모델링 – CNN 구조

<table>
<tr><td>1 x 1 Conv. 50
Batch-norm
ReLU</td><td>1 x 1 Conv. 80
Batch-norm
ReLU</td><td>1 x 1 Conv. 120
Batch-norm
ReLU</td><td rowspan="4">Flatten
Dense(50)
ReLU</td><td rowspan="4">Dense(20)
Softmax</td></tr>
<tr><td>3 x 3 Conv. 50
Batch-norm
ReLU</td><td>3 x 3 Conv. 80
Batch-norm
ReLU</td><td>3 x 3 Conv. 120
Batch-norm
ReLU</td></tr>
<tr><td>3 x 3 Conv. 50
Batch-norm
ReLU</td><td>3 x 3 Conv. 80
Batch-norm
ReLU</td><td>3 x 3 Conv. 120
Batch-norm
ReLU</td></tr>
<tr><td>2 x 2 MaxPool</td><td>2 x 2 MaxPool</td><td>2 x 2 MaxPool</td></tr>
</table>

Mission 3.

모델링 – 하이퍼파라미터, 이미지 크기 설정

```
batch_size = 128

# Training 수
epochs =10

# Weight 조절 parameter
LearningRate = 1e-3 # 0.001
Decay = 1e-6

img_width = 224
img_height = 224
```

batch_size : 128

epoch : 10

초기 learning rate : 1e-3

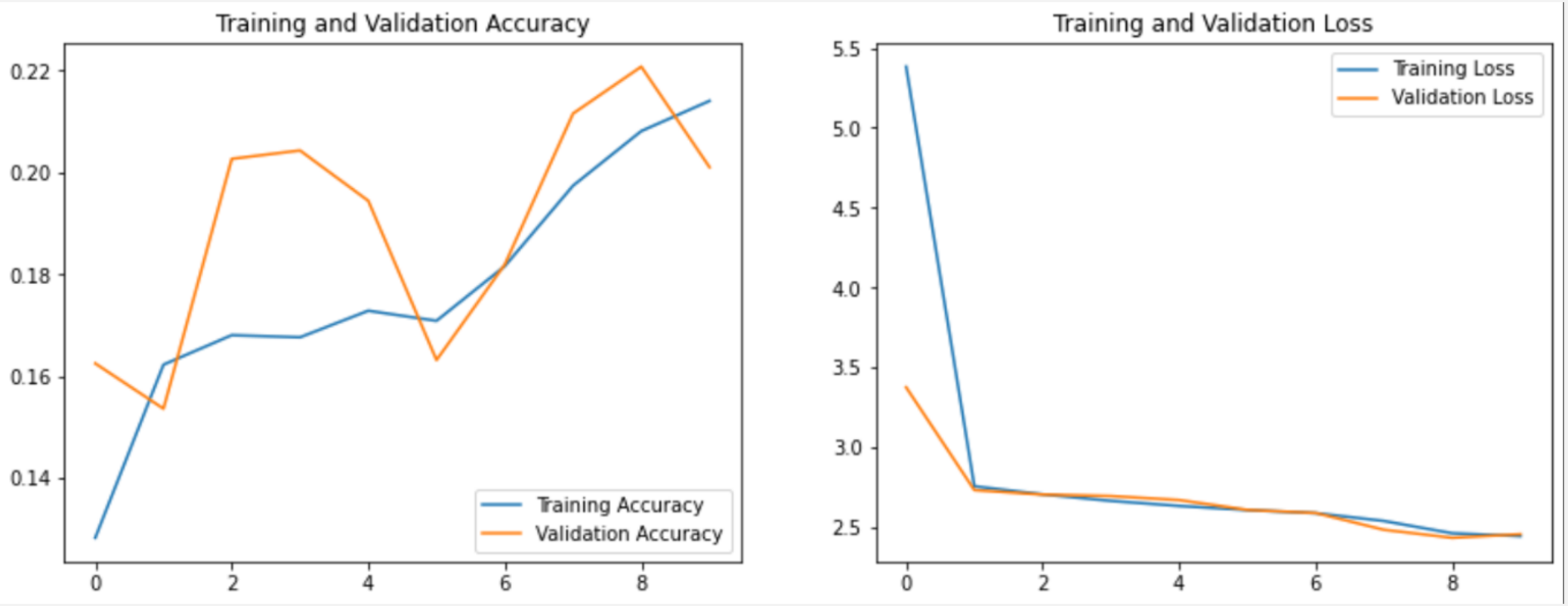
learning rate 변동폭 : 1e-6

img_width : 224

img_height : 224

Mission 3.

모델 학습



Validation set의 accuracy가 가장 높을 때를 최적의 모형으로 저장

Mission 3.

모델 평가 - F1 score

평균 F1-score는 0.135

Label	F1 Score	Label	F1 Score	Label	F1 Score	Label	F1 Score
L2_3	0.05	L2_21	0.00	L2_33	0.00	L2_44	0.00
L2_10	0.21	L2_24	0.02	L2_34	0.00	L2_45	0.00
L2_12	0.00	L2_25	0.00	L2_39	0.00	L2_46	0.22
L2_15	0.14	L2_27	0.00	L2_40	0.00	L2_50	0.02
L2_20	0.00	L2_30	0.00	L2_41	0.00	L2_52	0.00

감사합니다

2022 DATA CREATOR CAMP

과학기술정보통신부 NIA 한국지능정보사회진흥원