자료분석특론 Final Project 1

Wine Review

212STG21 이선경

INDEX

모델링 목표

와인 점수 (points) 최적 예측 모형 구하기

자료 설명

Points : (Numeric) 'WineEnthusiast'가 매긴 와인의 점수 (scale : 1-100)

Country : (Categorical) 와인이 생산된 국가 (43개국)

Description : (Categorical) 와인에 대한 설명

Designation : (Categorical) 와인을 생산하는데 이용된 포도가 재배된 포도 농장

Price : (Numeric) 와인 한 병의 가격

Province : (Categorical) 와인이 생산된 지역

Region_1 : (Categorical) 와인 포도 재배 지역 1

Region_2 : (Categorical) 와인 포도 재배 지역 2

Taster_name : (Categorical) 시음자 이름

Taster_twitter_handle : (Categorical) 시음자의 트위터 계정 이름

Title : (Categorical) 와인 이름

Variety : (Categorical) 와인의 포도 품종

Winery : (Categorical) 와이너리 이름

Description, Taster_twitter_handle 변수는 모두 unique한 변수이므로 제거하고 시작

MISSING 처리

변수별 Missing 개수

country	63
designation	37465
points	0
price	8996
province	63
region_1	21247
region_2	79460
taster_name	26244
title	0
variety	1
winery	0

dtype: int64

COLUMN제거

Missing 이 대부분인 'region_2' 변수 제거

'Unknown'으로 대체

'region_1'의 count

Napa Valley	4480
Columbia Valley (WA)	4124
Russian River Valley	3091
California	2629
Paso Robles	2350
...	
Sonoma County-Santa Barbara County	1
Mendocino-Amador	1
Tarantino	1
Saint-Chinian-Roquebrun	1
Vin de Pays de l'Atlantique	1

Name: region_1, Length: 1229, dtype: int64

'designation'의 count

Reserve	2006
Estate	1321
Reserva	1259
Riserva	698
Estate Grown	621
...	
Rico	1
Kapinas	1
Alpine Vineyard	1
De la Cava	1
Il Leccio	1

Name: designation, Lengt

'taster_name'의 count

Roger Voss	25512
Michael Schachner	15127
Kerin O'Keefe	10776
Virginie Boone	9537
Paul Gregutt	9531
Matt Kettmann	6332
Joe Czerwinski	5145
Sean P. Sullivan	4966
Anna Lee C. Iijima	4415
Jim Gordon	4177
Anne Krebiehl MW	3676
Lauren Buzzeo	1832
Susan Kostrzewa	1080
Mike DeSimone	502
Jeff Jenssen	469
Alexander Peartree	415
Carrie Dykes	139
Fiona Adams	27
Christina Pickard	6

Name: taster_name, dtype: int64

모두 데이터의 과반 이상을 차지하는 범주가 없으므로 NaN을 'Unknown'으로 대체

설명 변수 Missing처리 후 변수별 Missing 개수

```
country          0
designation      0
points           0
price         8992
province         0
region_1         0
taster_name      0
title            0
variety          0
winery           0
dtype: int64
```

반응변수 MISSING 처리

Points와 price의 상관계수가 0.4 이므로 Missing을 제외한 데이터를 이용해

Points = b0 + b1*price 로 회귀 직선을 구한 후 Points의 Missing을 predict 값으로 대체

	country	designation	points	price	province	region_1	taster_name	title
0	Portugal	Avidagos	87	15.000000	Douro	Unknown	Roger Voss	Quinta dos Avidagos 2011 Avidagos Red (Douro)
1	US	Unknown	87	14.000000	Oregon	Willamette Valley	Paul Gregutt	Rainstorm 2013 Pinot Gris (Willamette Valley)
2	US	Reserve Late Harvest	87	13.000000	Michigan	Lake Michigan Shore	Alexander Peartree	St. Julian 2013 Reserve Late Harvest Riesling ...
3	US	Vintner's Reserve Wild Child Block	87	65.000000	Oregon	Willamette Valley	Paul Gregutt	Sweet Cheeks 2012 Vintner's Reserve Wild Child...
4	Spain	Ars In Vitro	87	15.000000	Northern Spain	Navarra	Michael Schachner	Tandem 2011 Ars In Vitro Tempranillo-Merlot (N...
...	...	...	...	...	...	...	...	...
129902	Italy	Doga delle Clavule	86	21.786994	Tuscany	Morellino di Scansano	Unknown	Caparzo 2006 Doga delle Clavule (Morellino di...
129903	Portugal	Pacheca Superior	90	44.220254	Douro	Unknown	Roger Voss	Quinta da Pacheca 2013 Pacheca Superior Red (D...
129904	Portugal	Reserva	90	44.220254	Dão	Unknown	Roger Voss	Seacampo 2011 Reserva Red (Dão)
129905	Italy	Corte Menini	91	49.828569	Veneto	Soave Classico	Kerin O'Keefe	Le Mandolare 2015 Corte Menini (Soave Classico)
129906	France	Domaine Saint-Rémy Herrenweg	90	44.220254	Alsace	Alsace	Roger Voss	Domaine Ehrhart 2013 Domaine Saint-Rémy Herren...

129907 rows × 10 columns

YEAR 변수 생성

```
1 data_clean
```

	country	designation	points	price	province	region_1	taster_name	title	variety	winery	year
0	Portugal	Avidagos	87	15.000000	Douro	Unknown	Roger Voss	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos	2011
1	US	Unknown	87	14.000000	Oregon	Willamette Valley	Paul Gregutt	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm	2013
2	US	Reserve Late Harvest	87	13.000000	Michigan	Lake Michigan Shore	Alexander Peartree	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian	2013
3	US	Vintner's Reserve Wild Child Block	87	65.000000	Oregon	Willamette Valley	Paul Gregutt	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks	2012
4	Spain	Ars In Vitro	87	15.000000	Northern Spain	Navarra	Michael Schachner	Tandem 2011 Ars In Vitro Tempranillo-Merlot (N...	Tempranillo-Merlot	Tandem	2011
...	...	...	...	...	...	...	...	...	...	...	...
129902	Italy	Doga delle Clavule	86	21.786994	Tuscany	Morellino di Scansano	Unknown	Caparzo 2006 Doga delle Clavule (Morellino di...	Sangiovese	Caparzo	2006
129903	Portugal	Pacheca Superior	90	44.220254	Douro	Unknown	Roger Voss	Quinta da Pacheca 2013 Pacheca Superior Red (D...	Portuguese Red	Quinta da Pacheca	2013
129904	Portugal	Reserva	90	44.220254	Dão	Unknown	Roger Voss	Seacampo 2011 Reserva Red (Dão)	Portuguese Red	Seacampo	2011
129905	Italy	Corte Menini	91	49.828569	Veneto	Soave Classico	Kerin O'Keefe	Le Mandolare 2015 Corte Menini (Soave Classico)	Garganega	Le Mandolare	2015
129906	France	Domaine Saint-Rémy Herrenweg	90	44.220254	Alsace	Alsace	Roger Voss	Domaine Ehrhart 2013 Domaine Saint-Rémy Herren...	Gewürztraminer	Domaine Ehrhart	2013

129907 rows × 11 columns

- ‘title’ 변수에서 와인 제조년도를 뽑아 ‘year’ 변수 생성
- 'year' 변수의 Missing은 평균으로 대체
- ‘title’변수는 대부분이 unique이므로 제거

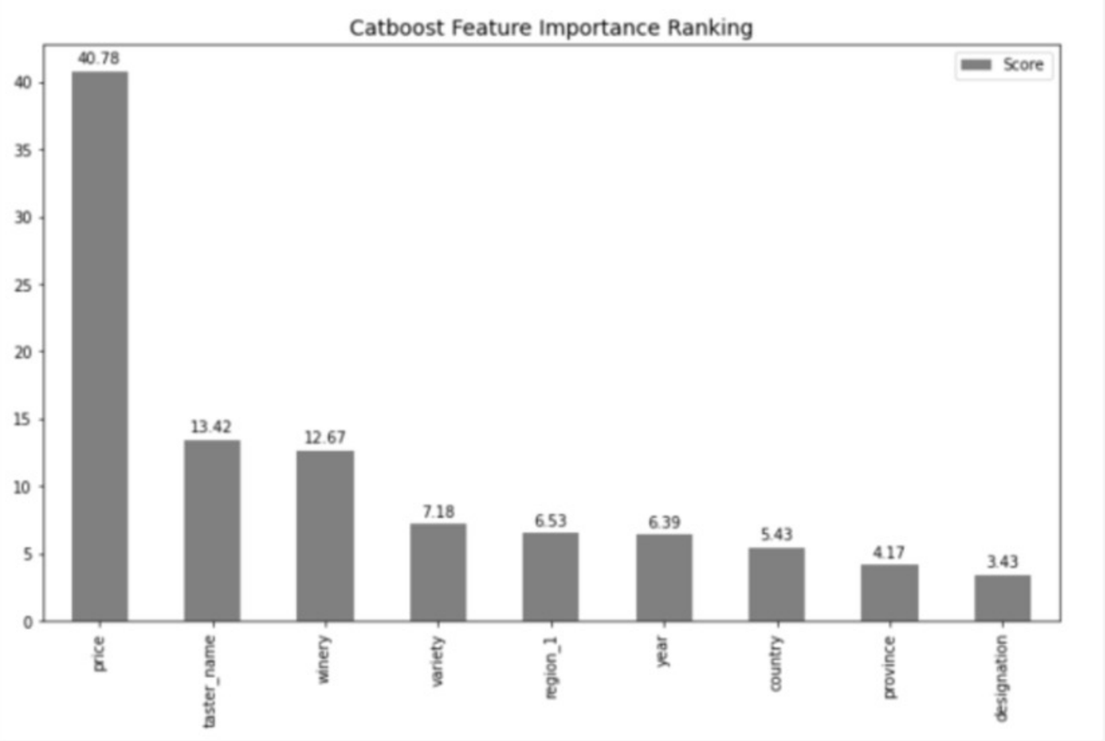
Scaling/Encoding

	country	designation	points	price	province	region_1	taster_name	variety	winery	year
0	31	2321	87	-0.509822	108	1094	15	447	12949	0.090608
1	40	35088	87	-0.534409	268	1218	14	433	13011	0.405242
2	40	27953	87	-0.558997	218	549	0	476	14380	0.405242
3	40	36441	87	0.719561	268	1218	14	437	14610	0.247925
4	37	1975	87	-0.509822	262	757	12	587	14695	0.090608
...	...	...	...	...	...	...	...	...	...	...
118775	22	10126	86	-0.342945	374	685	18	503	2207	-0.695978
118776	31	24593	90	0.208636	108	1094	15	447	12860	0.405242
118777	31	27732	90	0.208636	112	1094	15	447	13878	0.090608
118778	22	8151	91	0.346531	384	999	9	197	9867	0.719877
118779	15	10312	90	0.208636	11	21	15	209	5884	0.405242

118780 rows × 10 columns

- Numeric인 price와 year 변수는 Standard Scaling
- 나머지 Categorical 변수들은 Label Encoding

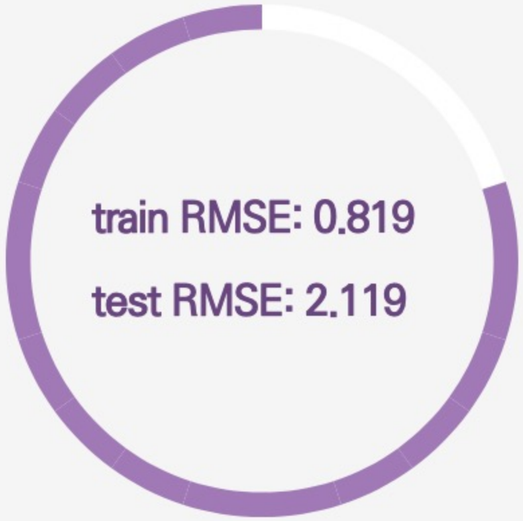
Feature Selection



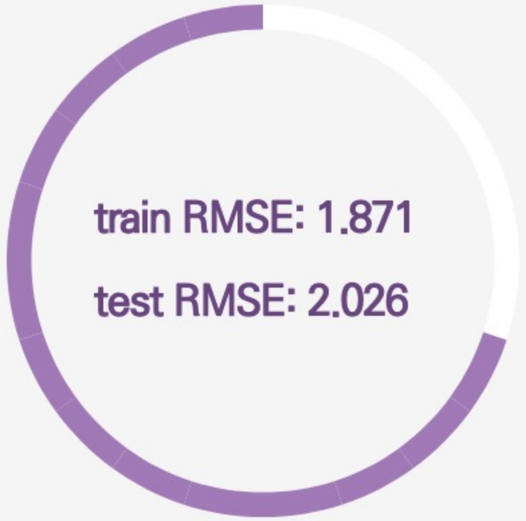
- 기본 Catboost 모형으로 fitting 후 그린 Feature Importance Plot
- 변수중요도가 낮은 'country', 'province', 'designation' 변수 제외

Basic Model

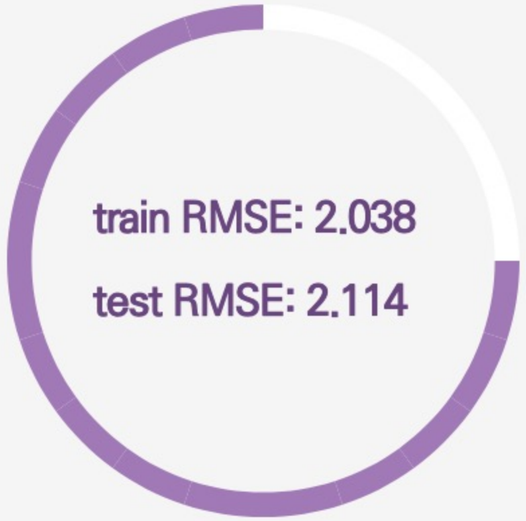
Random Forest



Light GBM



Catboost



Grid Search

Random Forest

Hyperparameters

'n_estimators' : [100 , 200, 500]
'max_features' : [2, 4, 6]

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(),
             param_grid={'max_features': [2, 4, 6],
                         'n_estimators': [100, 200, 500]},
             return_train_score=True, scoring='neg_mean_squared_error')
```

```
2.115132321950467 {'max_features': 2, 'n_estimators': 100}
2.1103518368003398 {'max_features': 2, 'n_estimators': 200}
2.1065148677562213 {'max_features': 2, 'n_estimators': 500}
2.1258432434610515 {'max_features': 4, 'n_estimators': 100}
2.1207801154440467 {'max_features': 4, 'n_estimators': 200}
2.118186815034524 {'max_features': 4, 'n_estimators': 500}
2.1381642245012666 {'max_features': 6, 'n_estimators': 100}
2.1318285857118595 {'max_features': 6, 'n_estimators': 200}
2.1288715214348404 {'max_features': 6, 'n_estimators': 500}
```

'max_features': 2, 'n_estimators': 500

train RMSE: 0.798
test RMSE: 2.094

Light GBM

Hyperparameters

'learning-rate' : [0.001, 0.01,0.05]
'max_depth' : [1, 7, 9, 13]

```
GridSearchCV(cv=5,
             estimator=LGBMRegressor(iterations=2000, loss_function='RMSE',
                                     random_seed=2021),
             param_grid={'learning-rate': [0.001, 0.01, 0.05],
                         'max_depth': [1, 7, 9, 13]},
             return_train_score=True, scoring='neg_mean_squared_error')
```

```
2.2452045996763403 {'learning-rate': 0.001, 'max_depth': 1}
2.053054818878503 {'learning-rate': 0.001, 'max_depth': 7}
2.0443920946333463 {'learning-rate': 0.001, 'max_depth': 9}
2.0371979907056654 {'learning-rate': 0.001, 'max_depth': 13}
2.2452045996763403 {'learning-rate': 0.01, 'max_depth': 1}
2.053054818878503 {'learning-rate': 0.01, 'max_depth': 7}
2.0443920946333463 {'learning-rate': 0.01, 'max_depth': 9}
2.0371979907056654 {'learning-rate': 0.01, 'max_depth': 13}
2.2452045996763403 {'learning-rate': 0.05, 'max_depth': 1}
2.053054818878503 {'learning-rate': 0.05, 'max_depth': 7}
2.0443920946333463 {'learning-rate': 0.05, 'max_depth': 9}
2.0371979907056654 {'learning-rate': 0.05, 'max_depth': 13}
```

'learning-rate': 0.001, 'max_depth': 13

train RMSE: 1.878
test RMSE: 2.029

최종 선택 모형

Catboost

Hyperparameters

'learning_rate' : [0.01, 0.1]
'iterations' : [500, 2000]

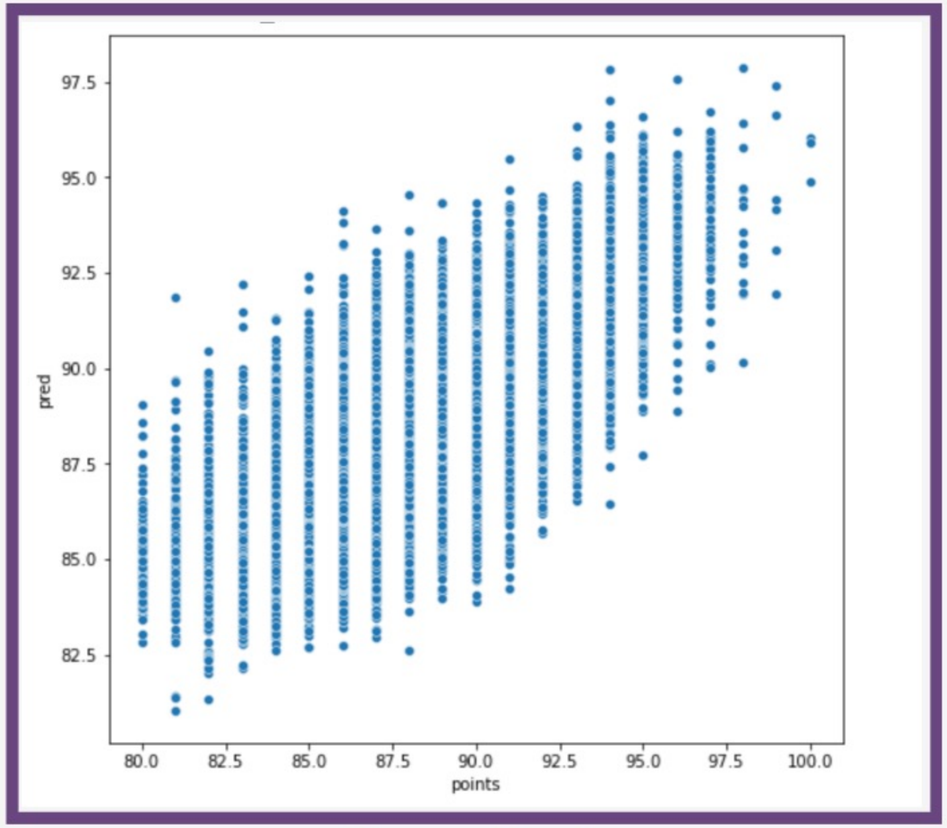
```
GridSearchCV(cv=5,
             estimator=<catboost.core.CatBoostRegressor object at 0x7f0e7bff92d0>,
             param_grid={'iterations': [500, 2000],
                         'learning_rate': [0.01, 0.1]},
             return_train_score=True, scoring='neg_mean_squared_error')
```

```
2.0989057971010983 {'iterations': 500, 'learning_rate': 0.01}
2.0070380753427877 {'iterations': 500, 'learning_rate': 0.1}
2.0306562760253573 {'iterations': 2000, 'learning_rate': 0.01}
1.9950572000648625 {'iterations': 2000, 'learning_rate': 0.1}
```

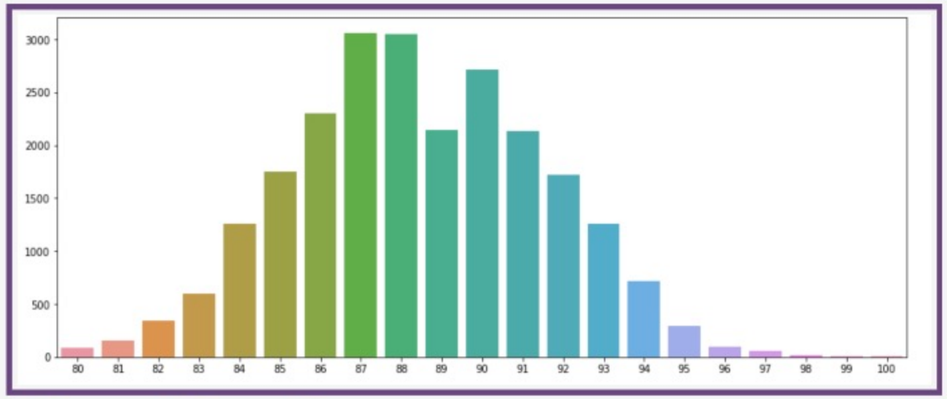
'iterations': 2000, 'learning_rate': 0.1

train RMSE: 1.780
test RMSE: 1.977

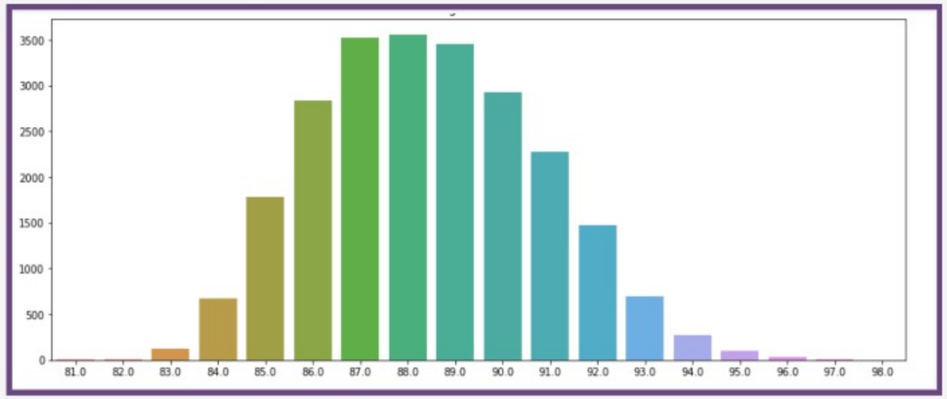
FINAL MODEL



True vs. Predict



Histogram of True points



Histogram of Predicted points

- Test data의 True points와 Predicted points의 분포를 보면 작은 값은 크게, 큰 값은 작게 예측한 경향이 보인다.

자료분석특론 Final Project 2

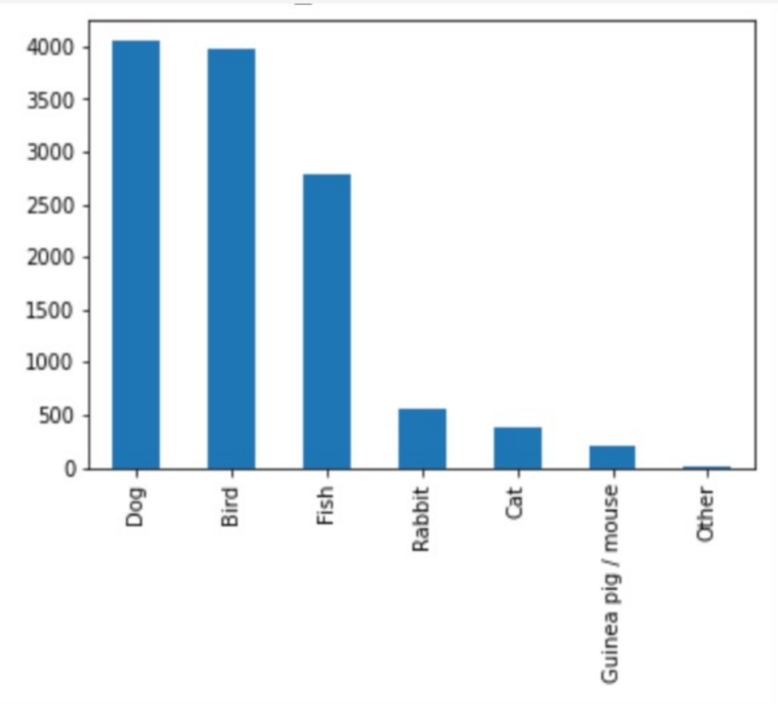
Animal Image

212STG21 이선경

INDEX

모델링 목표

동물 타입 (Animal_Type)
Image Classification 최적 CNN 모형 구하기
Class 수 = 7

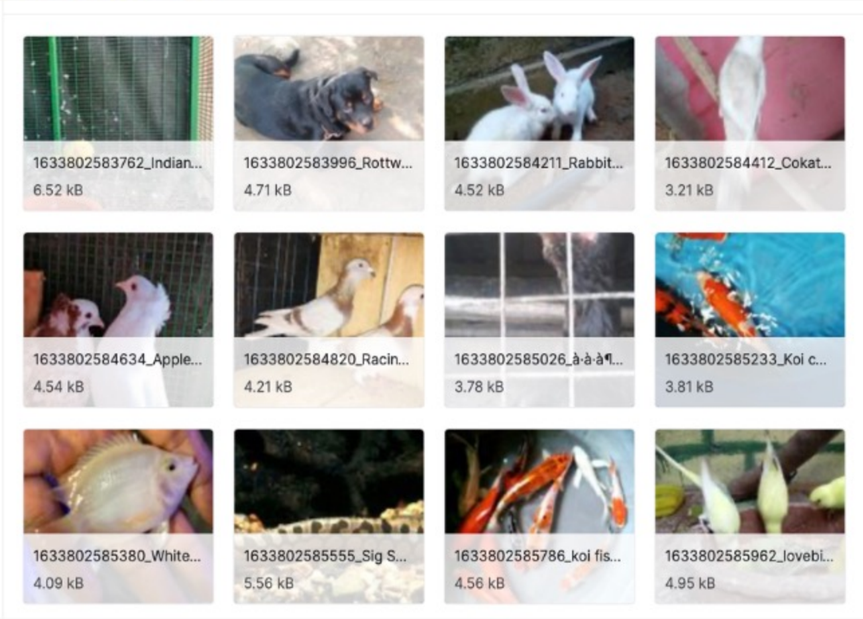


자료 설명

12만개의 라벨 처리한 동물 이미지 파일(.jpg) 데이터

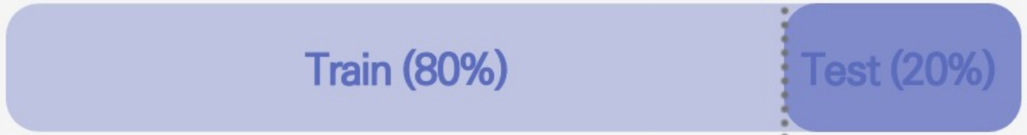
Annotated Label : 동물 세부종 라벨
Animal_Type : 동물 타입
Image_File : 동물 이미지 파일 경로

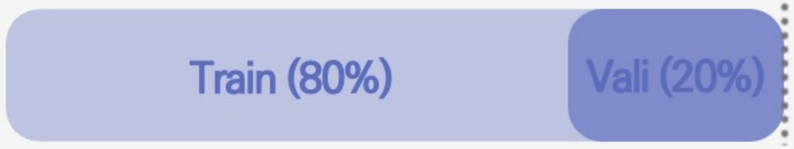
이미지 파일 예시)



Train – test Split

Train set 64 %
Validation Set 16 %
Test set 20%





Batch size =128

이미지 불러오기

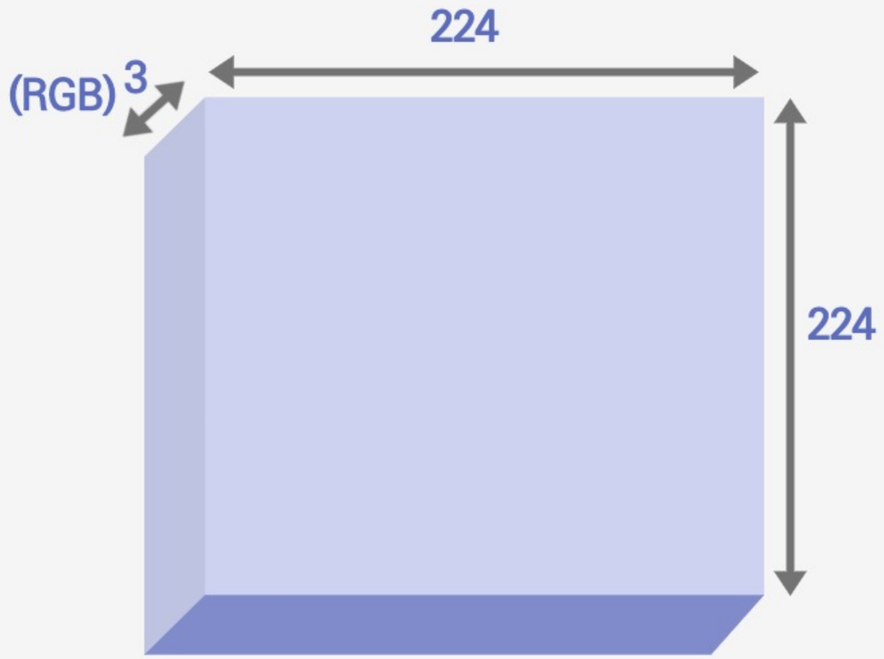
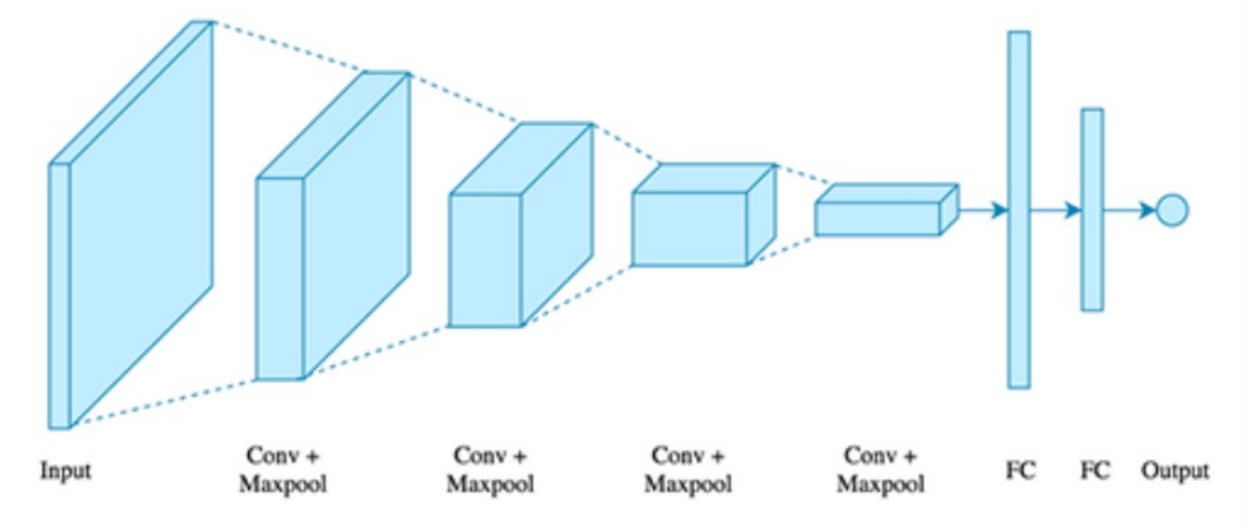
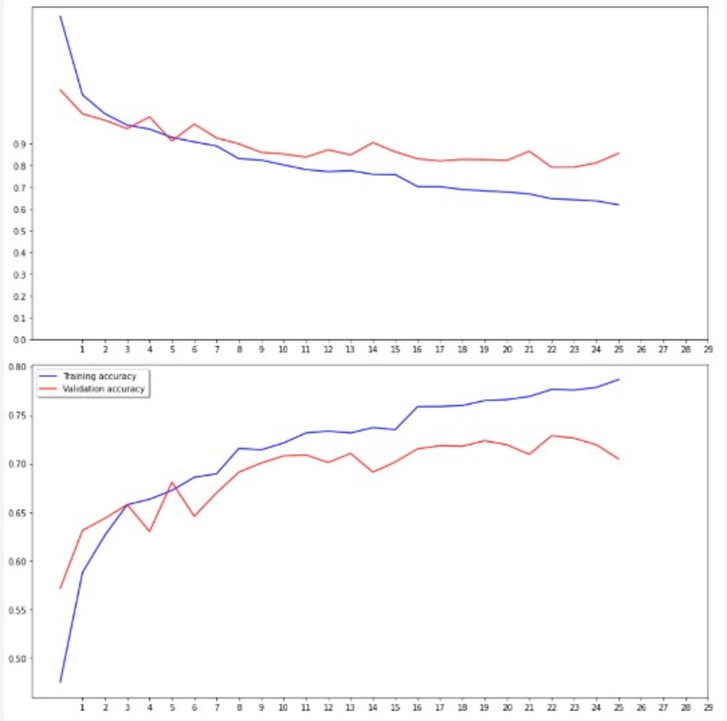


Image size = (224 , 224)
Image Channels = 3

CNN – 1

Conv2d (3 x 3, 32) layer – Max pooling(2 x 2) – Conv2d (3 x 3, 64) layer – Max pooling(2 x 2) – Conv2d (3 x 3, 128) layer – Max pooling(2 x 2) – Conv2d (3 x 3, 64) layer – Flatten – Fully Connected layer (512, relu) – Fully Connected layer (7, softmax)

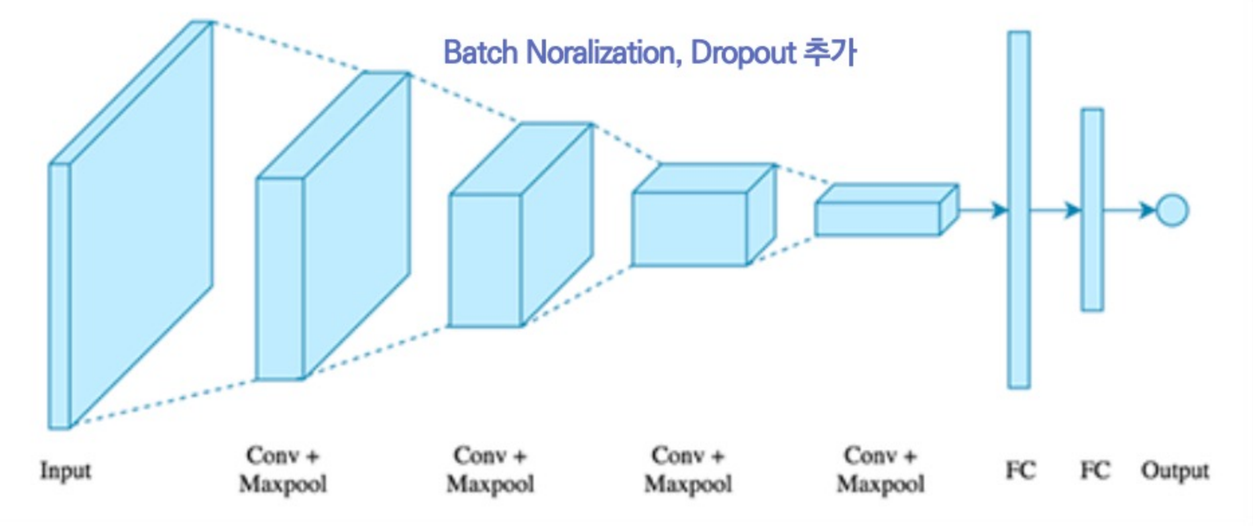


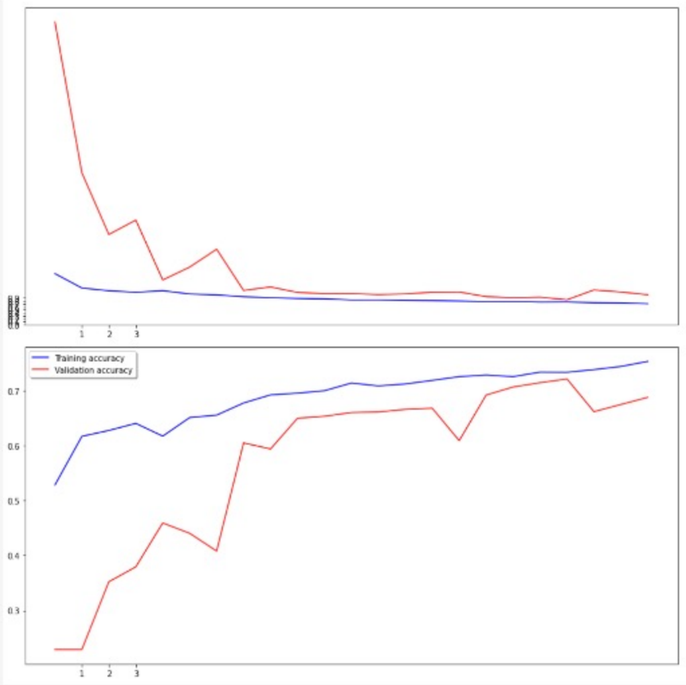


최종 epoch 26에서 val_loss: 0.8553 – val_accuracy: 0.7048

CNN – 2

Conv2d (3 x 3, 32) layer – BN –Max pooling(2 x 2) – Dropout – Conv2d (3 x 3, 64) layer – BN – Max pooling(2 x 2) – Dropout – Conv2d (3 x 3, 128) layer – BN – Max pooling(2 x 2) – Dropout – Conv2d (3 x 3, 64) layer – Flatten – Fully Connected layer (512, relu) – BN – Dropout – Fully Connected layer (7, softmax)

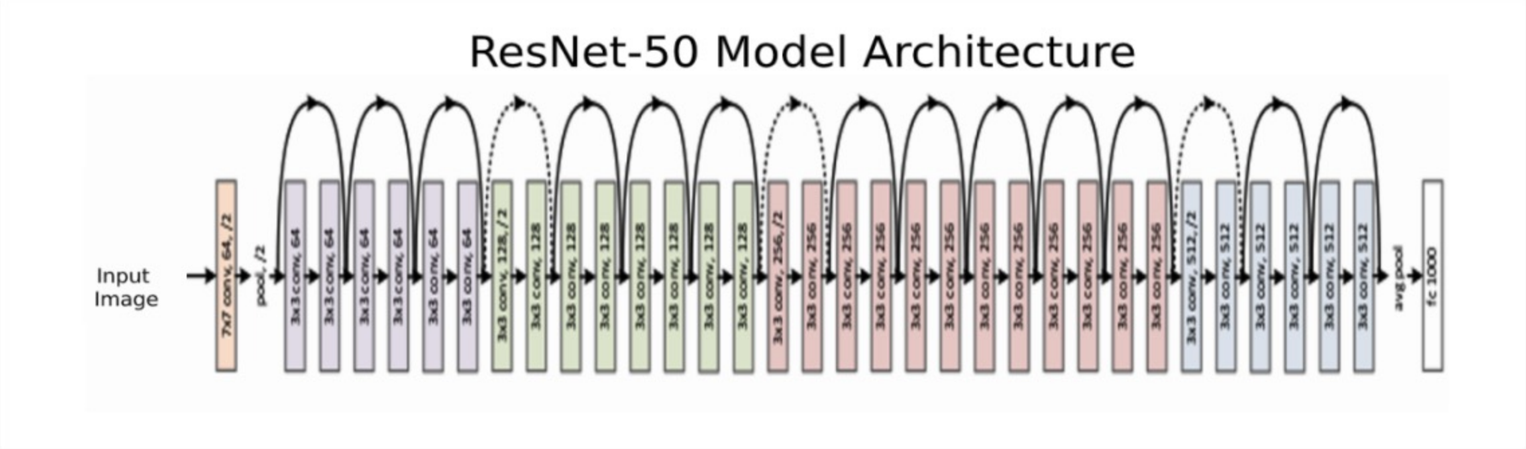


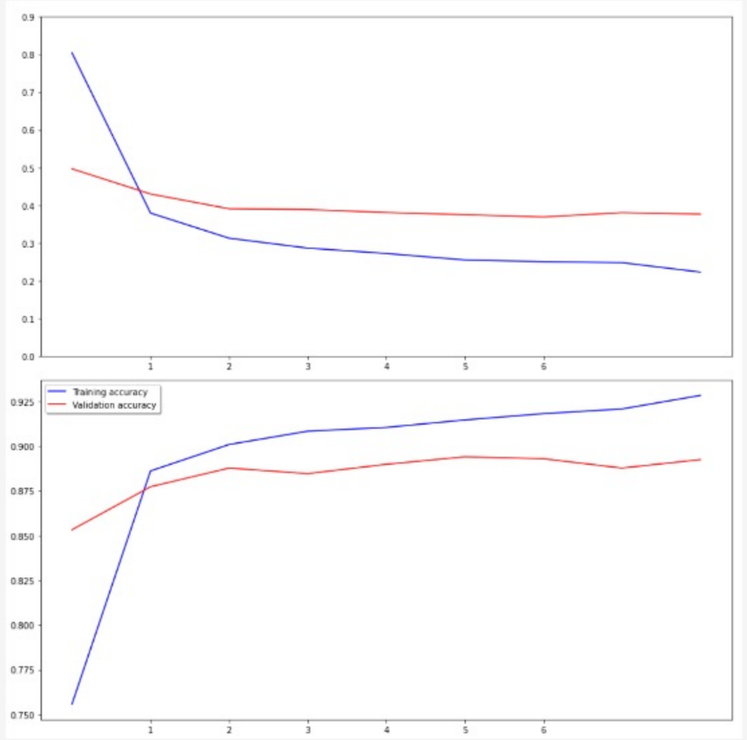


최종 epoch 23에서 val_loss: 1.0118 – val_accuracy: 0.6881

RESNET 50

Resnet50 – Fully Connected layer (7, softmax)

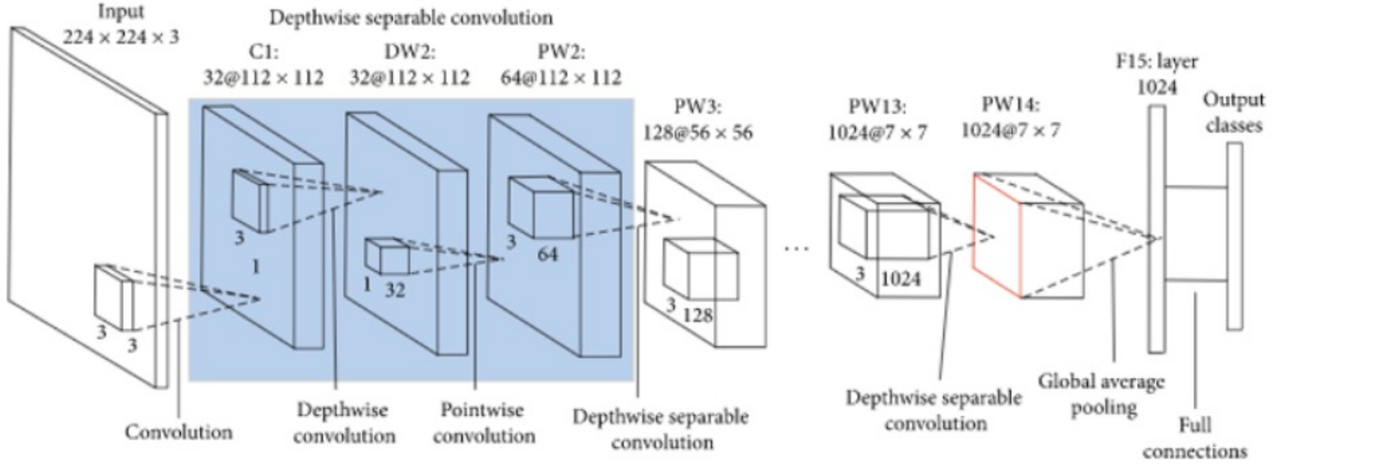




최종 epoch 8에서 val_loss: 1.3315 – val_accuracy: 0.7821

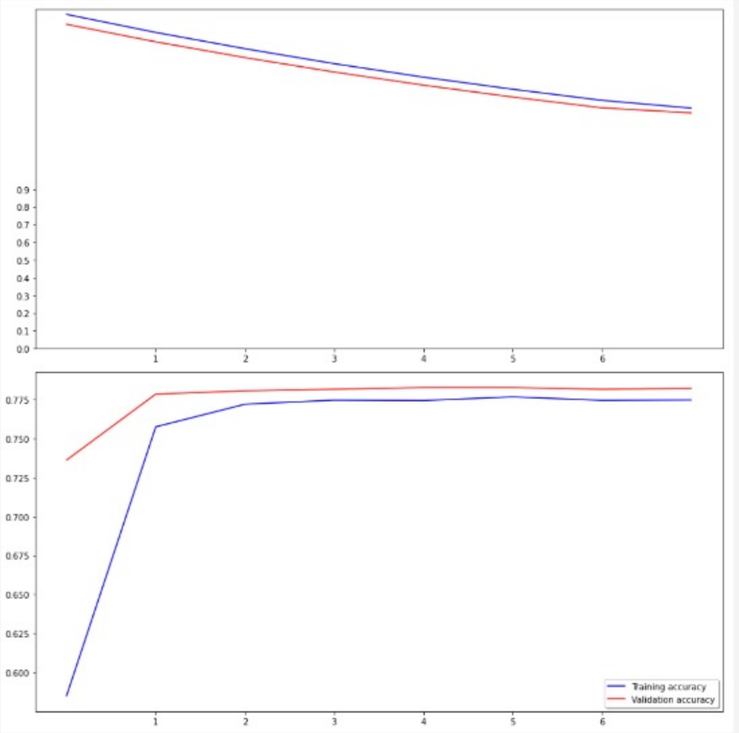
MOBILENET.V2

MOBILENET.V2 – Fully Connected layer (7, softmax)



MobileNet-V2 Architecture

Chiung-Yu Chen



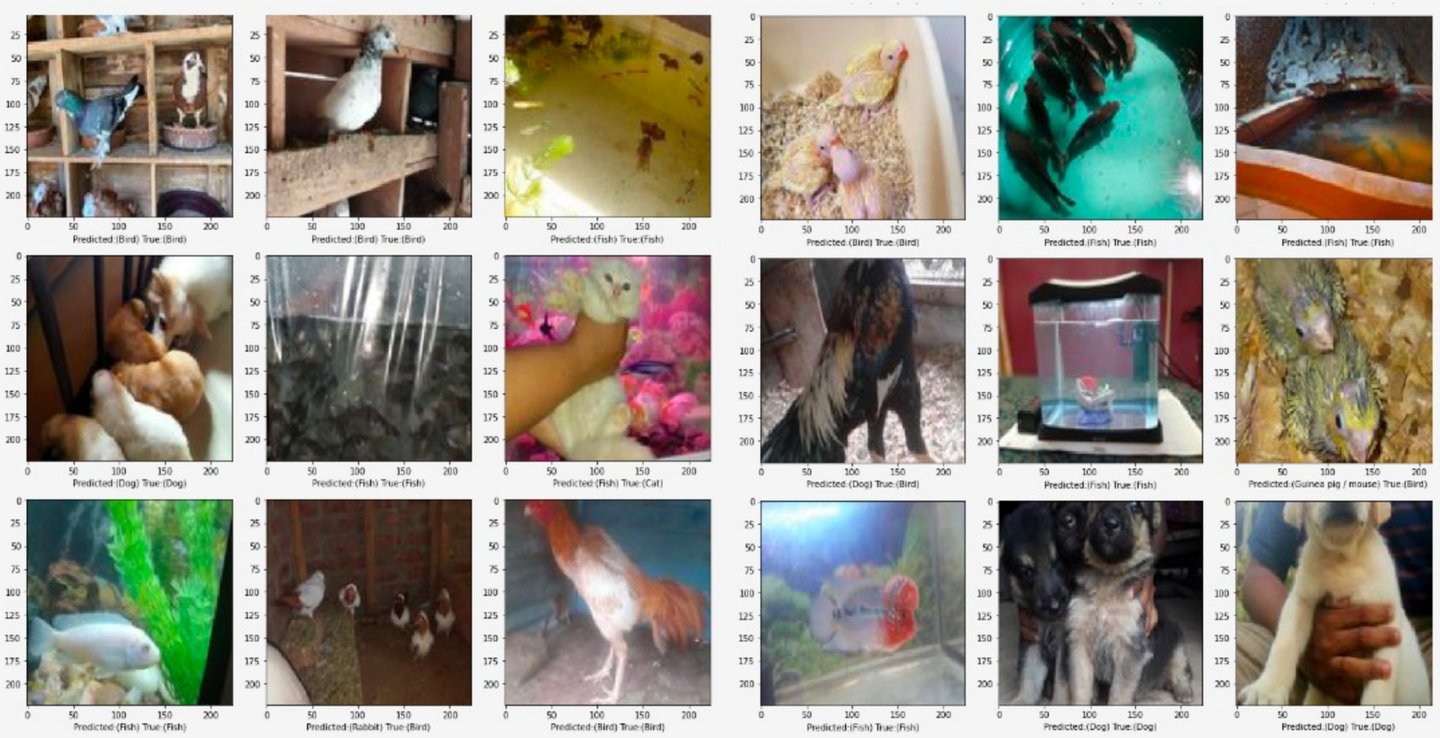
최종 epoch 9에서 val_loss: 0.3768 – val_accuracy: 0.8924

Test Accuracy

CNN – 1	CNN – 2	RESNET	MOBILENET.V2
0.720	0.701	0.799	0.903

MOBILENET.V2

– 최종 모형으로 구한 Test set Confusion Matrix



– 최종 모형 Test set Predict / True 라벨

자료분석특론 Final Project 2

감사합니다.

212STG21 이선경