경북소프트웨어고등학교

전국 청소년 인공지능 경진대회

조장: 우영웅
조원: 이재경, 김민규

CONTENTS

전국 청소년 인공지능 경진대회

CHAPTER.1

데이터 수집

AI 허브에서 라벨링 된 146,750 데이터를 가져온 후 30 * 30이하의 이미지 분류 제거 및 중복,오류 이미지 재검출

① AI Hub

- [ ] [라벨]c_train_1280_720_daylight_15.tar
- [ ] [원천]c_train_1920_1200_night_1.tar
- [x] 표지판코드분류crop데이터2.tar
- [x] 표지판코드분류crop데이터1.tar
- [ ] [원천]c_train_1280_720_night_1.tar
- [ ] [원천]d_train_1920_1080_night_1.tar
- [ ] [원천]c_train_1280_720_daylight_18.tar

②

```python
import os
from PIL import Image
import pandas as pd
import re
import shutil
# 1125 * 2000 기준

## 폴더 설정
home =  "D:/교통안전 경진대회/이미지"
# files = os.listdir("D:/교통안전 경진대회/테스트")
data = []
for (path, dir, files) in os.walk(home):
    # print(path, dir, files)
    if len(files) != 0 :
        for filename in files:
            data.append(path + '/' + filename)

# print(files[0
## 디렉토리 설정
# dir = os.getcwd()
# print(dir)
## 폴더 안 파일 사이즈 확인 for문
# print(files[0])
for index, i in enumerate(data):
    a = Image.open(i)
    w,h = a.size
    if w > 30 or h > 30:
        if index % 50 == 0:
            print('.', end='')
    else:
        a.close()
        os.remove(i)
```

③

데이터 평균화

DOWN SAMPLING

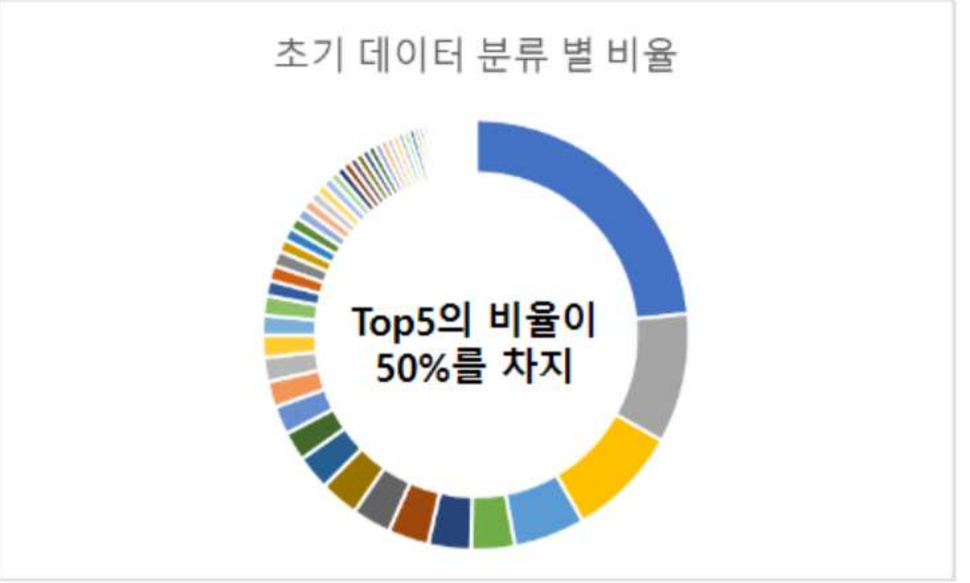
Data Imbalance 문제란?

Data Imbalance 문제 :특정 분류에 데이터가 편중되면 AI 모델의 학습이 제대로 진행되지 않는 문제

Imbalance 대책

- 적은 데이터를 늘린다(Up sampling)
  - : 추가 데이터 수집
  - : GAN 등 이미지 생성
- 많은 데이터를 줄인다(Down sampling)
  - : 학습 데이터 감소
  - : 방법론이 단순

AI Hub 데이터의 각 분류 별 비율



데이터 평균화

DOWN SAMPLING

DOWN SAMPLING 전

DOWN SAMPLING 후

DOWN SAMPLING

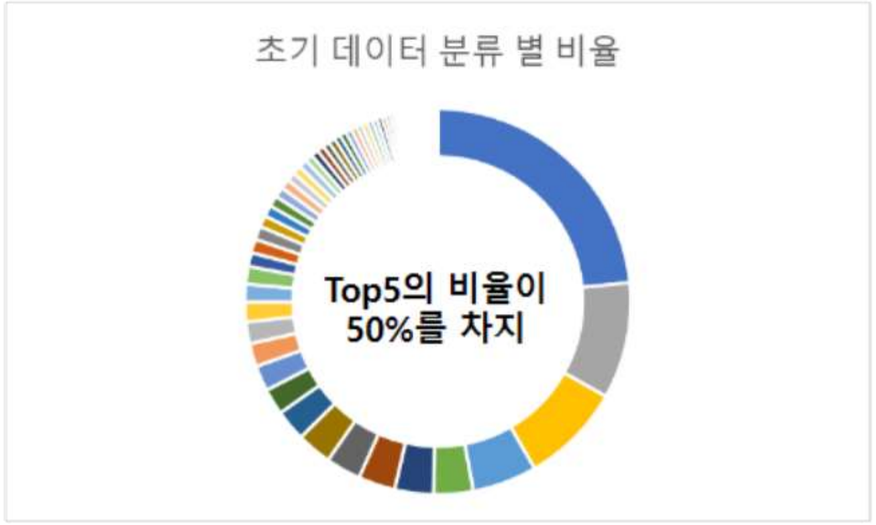
데이터 불균형 문제로 AI학습에 적당하지 못하여 1,000장 이상의 이미지를 Down sampling하여 800개 이하로 맞추어주었다.

데이터 평균화

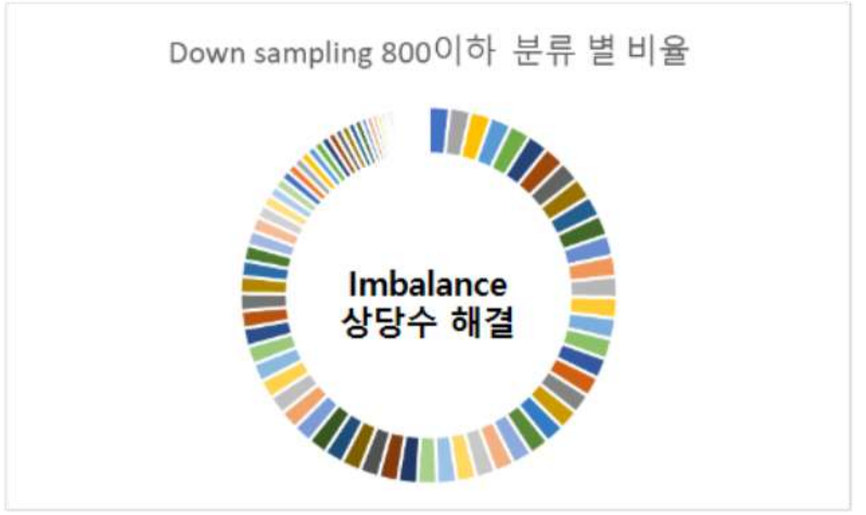
DOWN SAMPLING

데이터 불균형으로 인하여 T표지판을 많이 학습한 십자형으로
인식해버리는 형상을 방지하기 위해 평균화를 시킨다.

AI Hub 데이터의 각 분류 별 비율

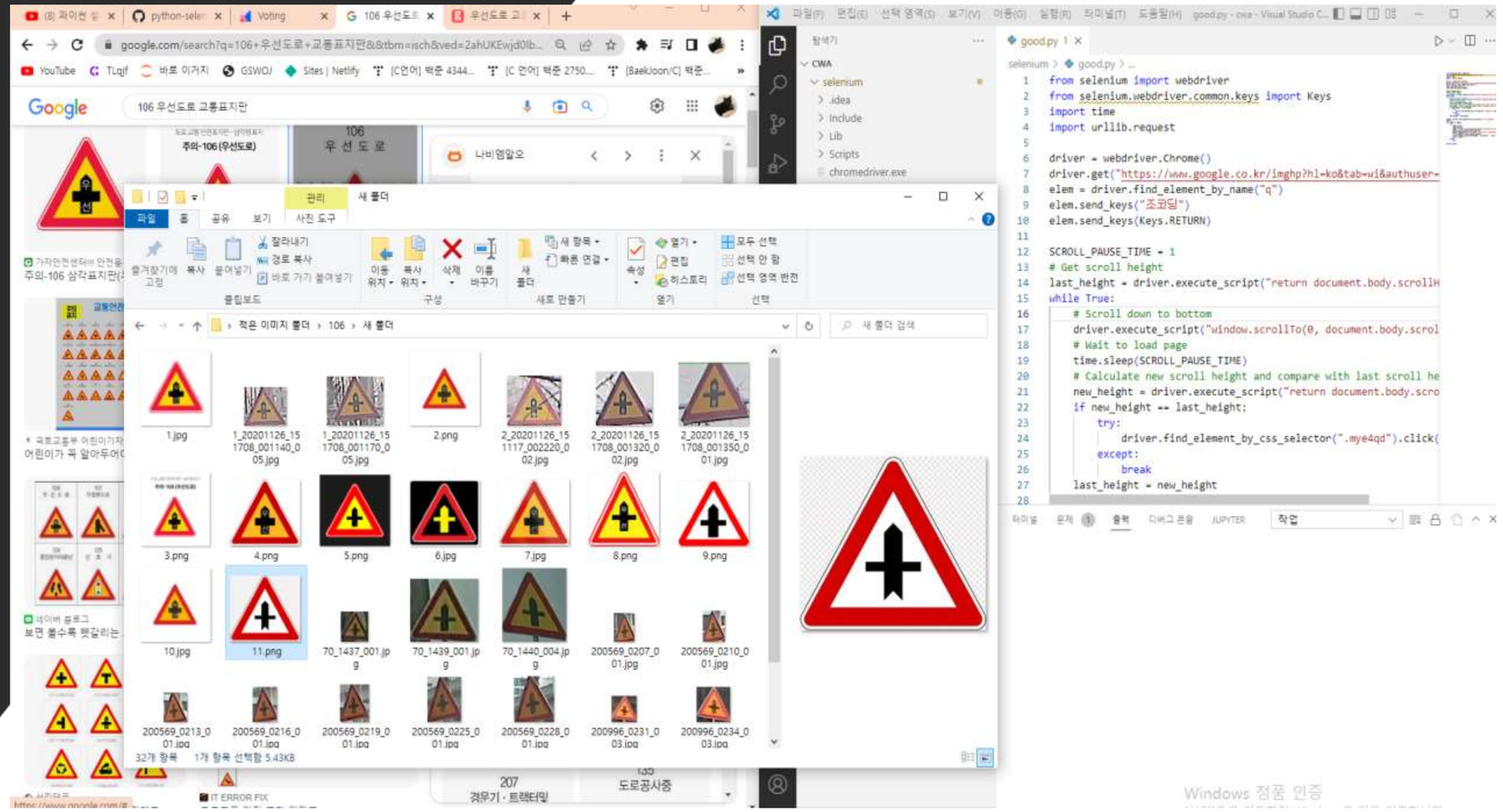


오분류 수정, 데이터 크롤링 추가 후 Down sampling



데이터 평균화

WEB CRAWLING

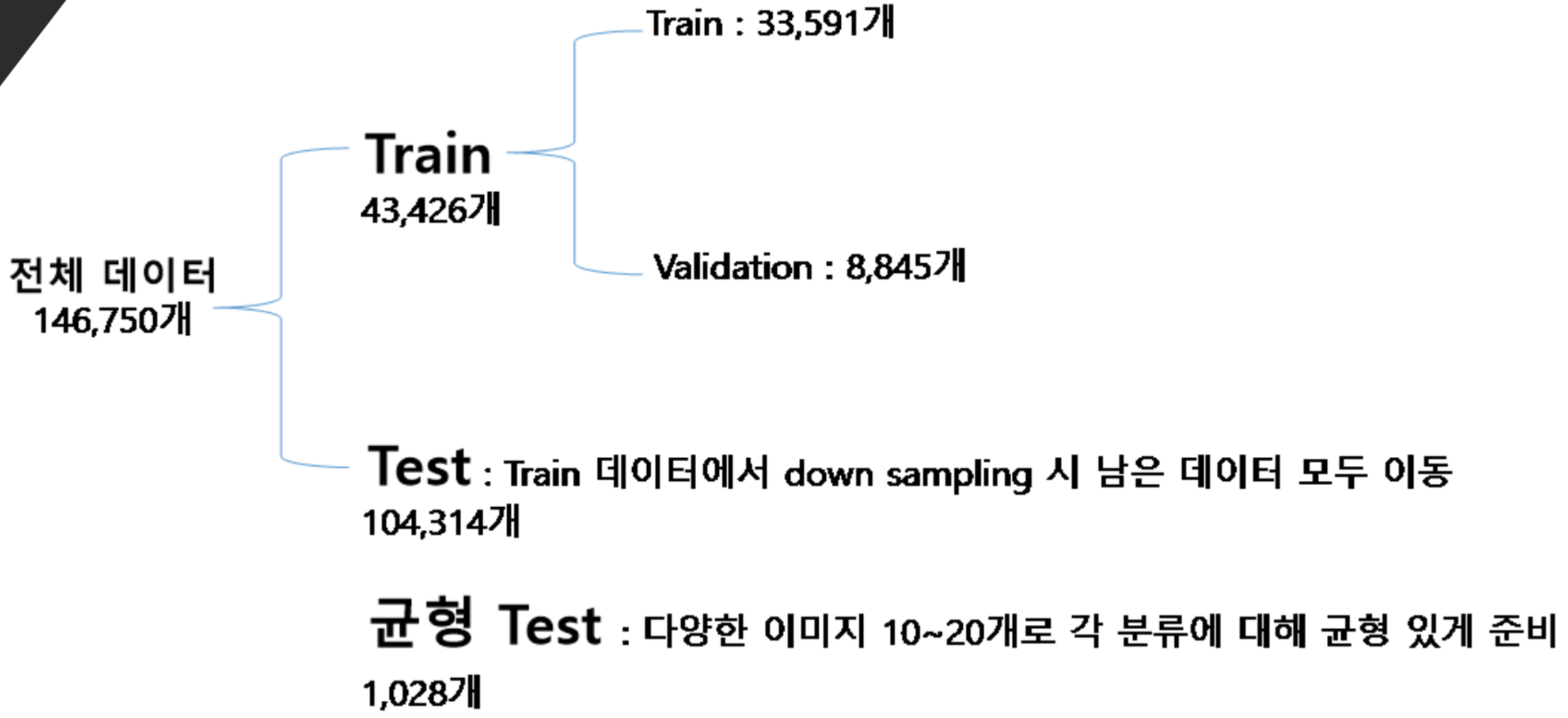


WEB CRAWLING을 사용하여 데이터 수집

WEB CRAWLING

표지판 폴더마다 양이 달라 적은 폴더는 데이터가 없는 경우가 있어 web crawling프로그램을 사용하여 부족한 이미지파일을 수집하였다.

Data 준비 – 최종 학습 및 평가 데이터 준비 완료



PROJECT

CHAPTER.2
모델 학습

- Model - Base 모델 보다 큰 모델 적용
- 평가 결과 Down sampling에 따른 성능 변화
- 평가 결과 Batch size에 따른 모델 성능 변화

04 Model - Base 모델 보다 큰 모델 적용

-. Base(6 분류) 보다 많은 분류(100 분류)를 해야하므로 좀 더 큰 모델(Trainable parameter가 큰)을 이용

1. Base 보다 Conv2D layer를 Max pooling 전에 추가
2. Pooling을 4번 적용하여 Flatten 후 모델 Shape이 Base 보다 5배 커짐

Base 코드 AI 모델 구조

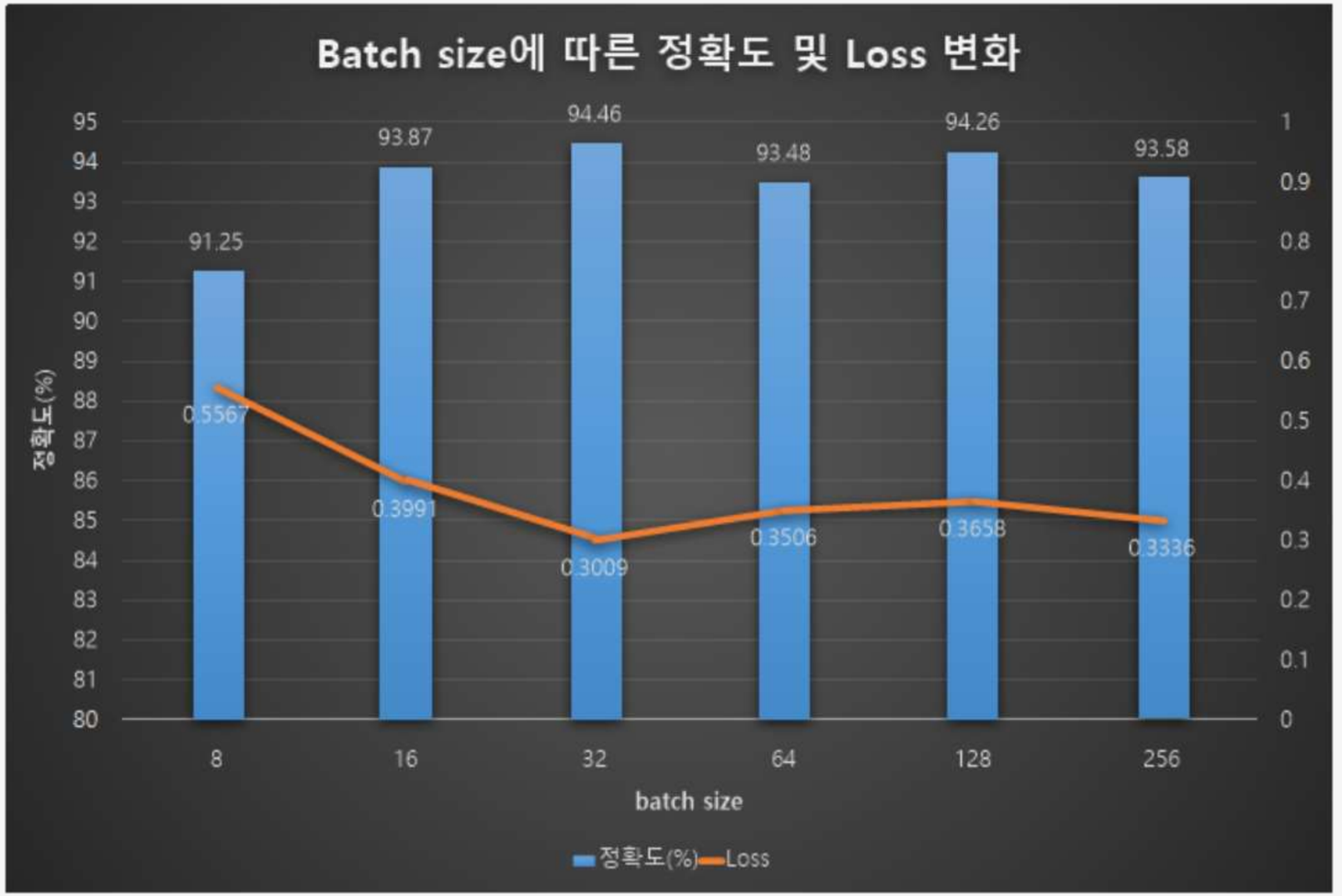
Layer	Kernel	Activation	Output Shape
Conv2D	3X3	relu	148X148X64
Max Pooling	2X2	-	74X74X64
Conv2D	3X3	relu	72X72X128
Max Pooling	2X2	-	36X36X128
Conv2D	3X3	relu	34X34X128
Max Pooling	2X2	-	17X17X128
Conv2D	3X3	relu	15X15X128
Max Pooling	2X2	-	7X7X128
Conv2D	3X3	relu	5X5X128
Max Pooling	2X2	-	2X2X128
Flatten	-	-	512
Dropout	-	-	512
Dense	-	relu	512
Dense	-	relu	128
Dense	-	relu	32
Dense	-	softmax	6
전체 Trainable parameters	**851,046**		

개선 적용 AI 모델 구조

Layer	Kernel	Activation	Output Shape
Conv2D	2X2	relu	149X149X64
Conv2D	3X3	relu	147X147X64
Max Pooling	2X2	-	73X73X64
Conv2D	2X2	relu	72X72X96
Conv2D	3X3	relu	70X70X96
Max Pooling	2X2	-	35X35X96
Conv2D	2X2	relu	34X34X128
Conv2D	3X3	relu	32X32X128
Max Pooling	2X2	-	16X16X128
Conv2D	2X2	relu	15X15X160
Conv2D	3X3	relu	13X13X160
Max Pooling	2X2	-	4X4X160
Flatten	-	-	2560
Dropout	-	-	2560
Dense	-	relu	512
Dense	-	relu	128
Dense	-	softmax	100
전체 Trainable parameters	**2,044,643**		

평가 결과 Batch size에 따른 모델 성능 변화

-. 균형 Test에 대해서 정확도 및 Loss 값 확인 시 Batch size 32에서 최대 정확도와 최소 Loss값을 가짐
본 발표에서 적용한 모델과 데이터에 대해서 Batch size 32 적용 데이터가 최적



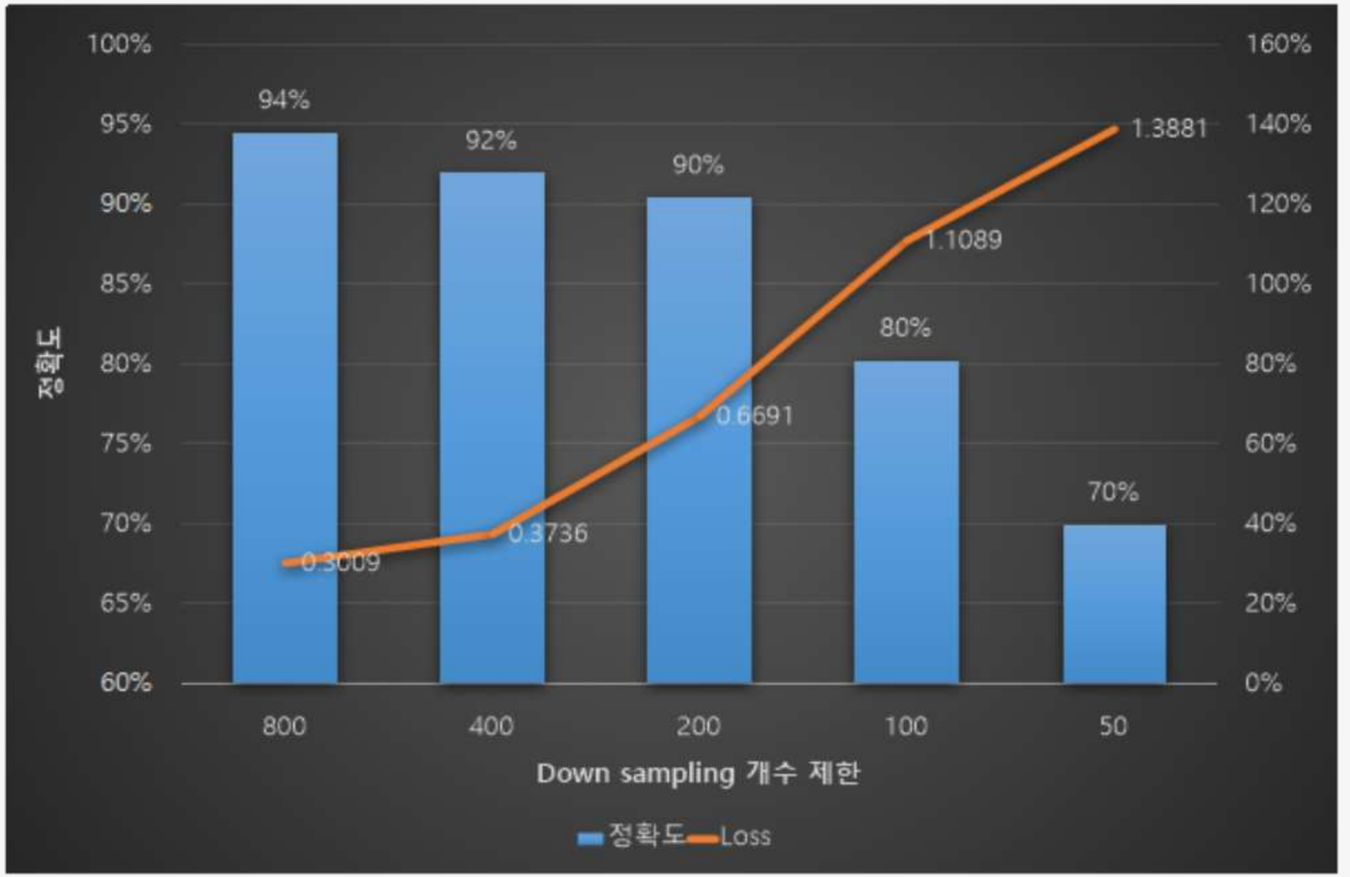
평가 결과 Down sampling에 따른 성능 변화

-. Batch size 32 조건, Down sampling 조건 8000이하, 4000이하, 2000이하, 1000이하, 500이하로 평가

-. Data 양과 Date Imbalance 문제에서 어느 것이 더 유리한가 탐색

Data 양이 줄어들면서 성능 저하가 나타나다, 특정 이하로 급격히 저하됨

: Down sampling보다는 Up sampling이 유리하다고 생각됨



CHAPTER.3
아쉬웠던 점

- 데이터 수집에 관해 아쉬웠던 점
- 모델 학습에 관해 아쉬웠던 점
- 느낌점과 아쉬웠던 점

아쉬웠던 점

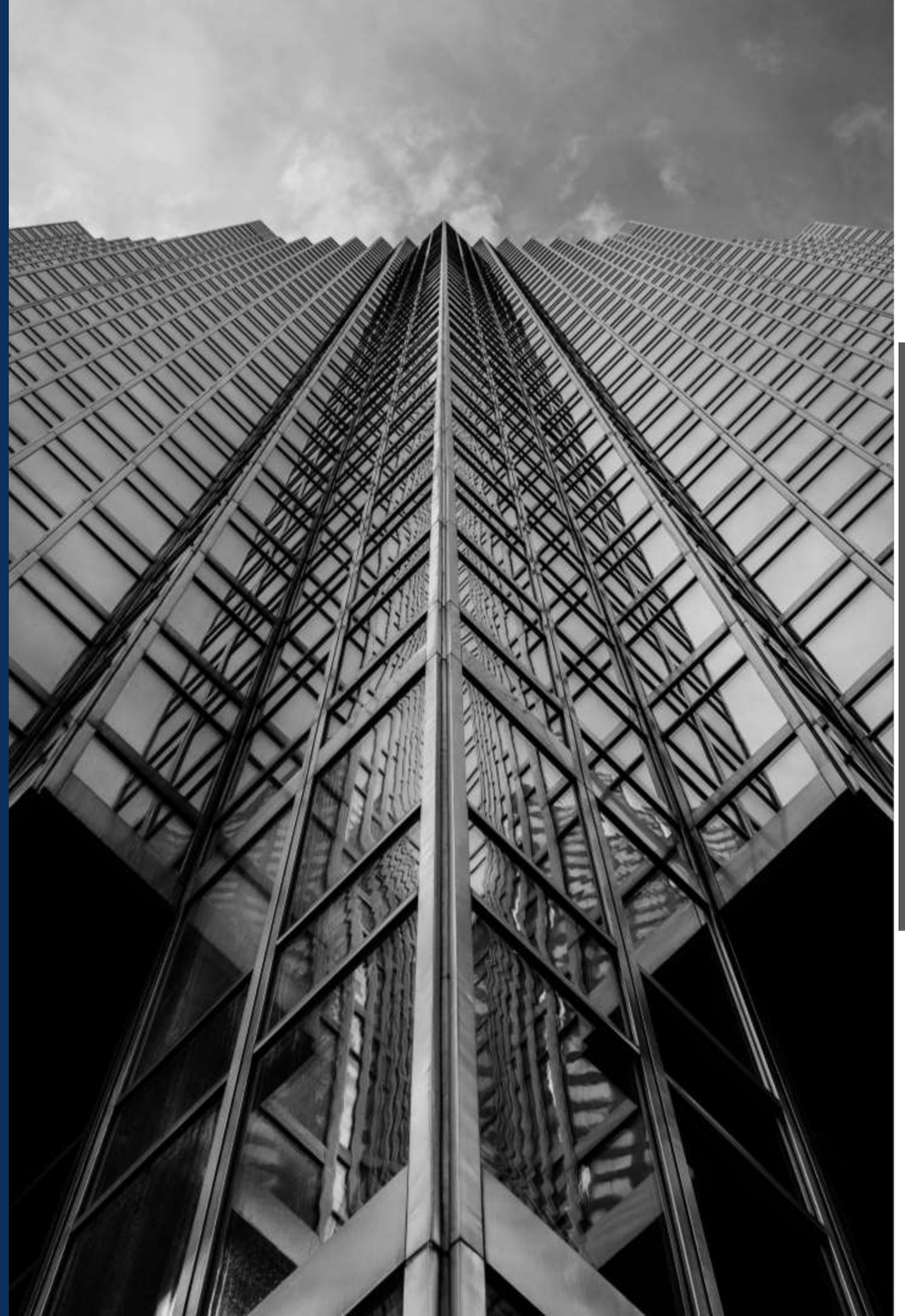
데이터 수집

web crawling 과정에서 구글에 관련 이미지가 부족하여 아쉬웠다.

모델 학습

모델을 학습시키기 위한 소모 시간이 길어 다양한 조건에서 최적화 진행을 하지 못해 아쉬웠다.

데이터 수집부터 모델 평가까지 전체 과정을 진행하면서 대회라는 것에서 다양한 생각을 하며 진행 할 수 있었습니다.
데이터 오분류 및 수집은 많은 노동이 필요한 작업으로 모델 성능에 가장 중요한 작업이여서 가장 많은 시간을 할애하여 진행하였습니다.



THANK YOU

조장: 우영웅
조원: 이재경, 김민규

경북소프트웨어고등학교