

# [CV-03] Data Centric

## 글자 검출 프로젝트

못해도 GEN참아

팀원 : 김준영, 신우진, 천지은, 전형우, 김승기

# INDEX

---

1. 팀 목표
2. EDA
3. Data Relabeling
4. 실험
5. 앙상블
6. 협업 방식

---

# 1. 팀 목표

---

이번 프로젝트 팀의 목표

## “데이터에 집중, 협업”

Annotation 가이드 작성, 재라벨링, 라벨링 검수

적극적인 의견 공유, 비판적 의견 제시하기

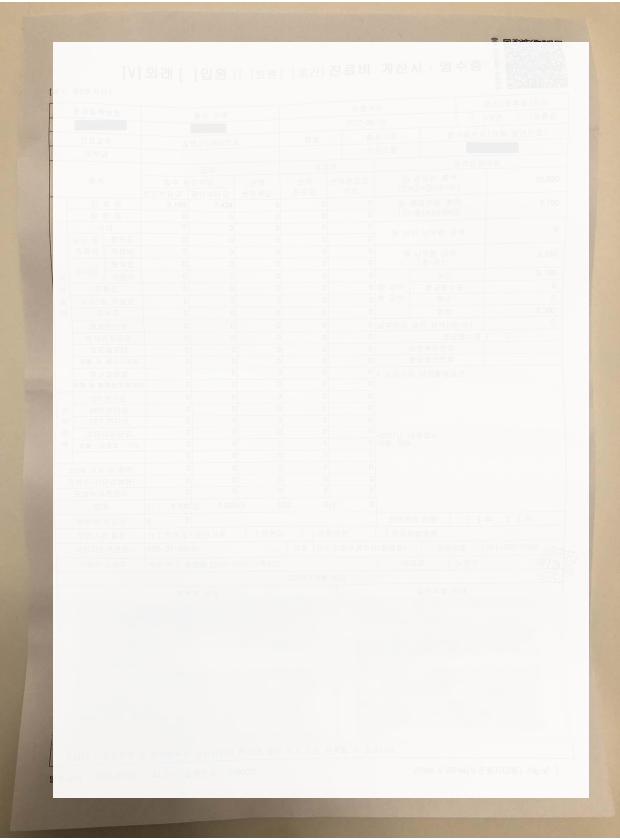
---

# 2. EDA

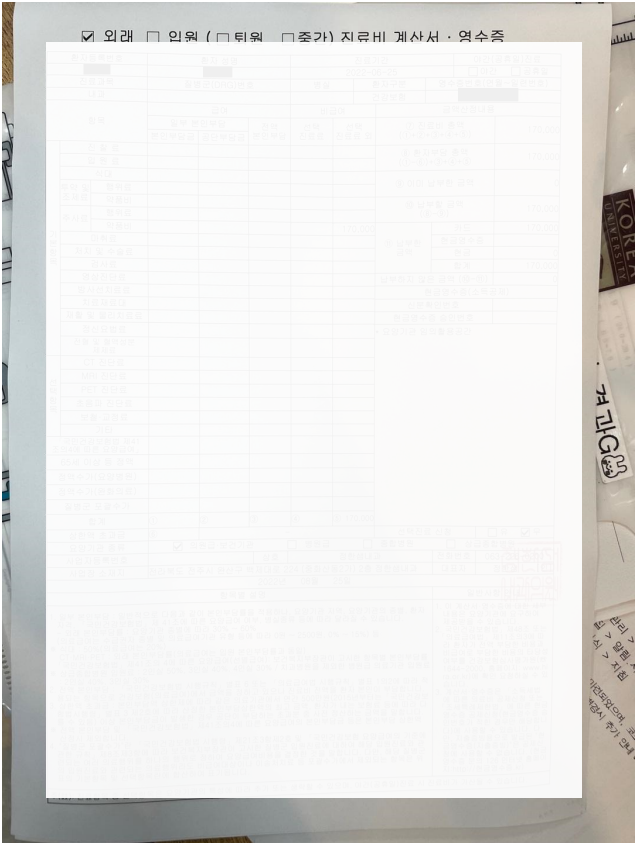
---

EDA

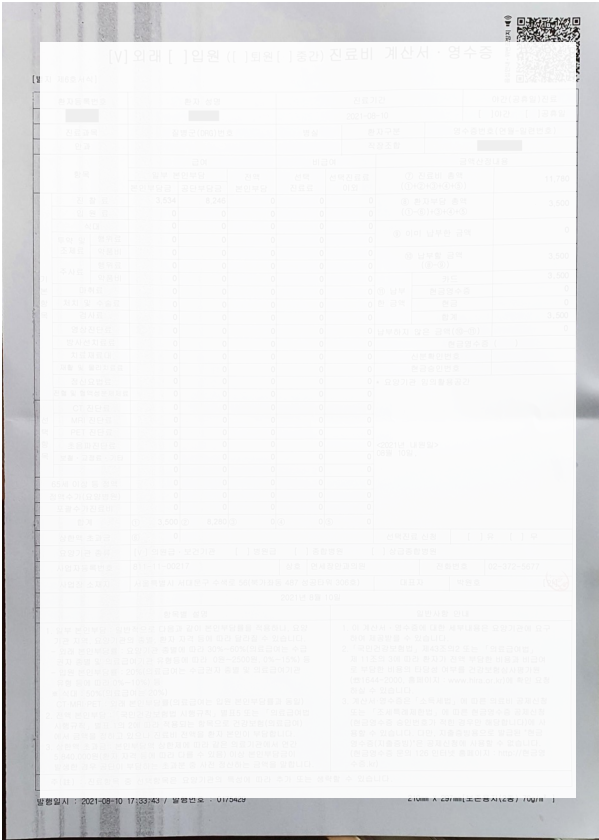
# 01. Train Data 파악



구겨진 이미지



배경이 있는 이미지



종이에 노이즈가 있는 이미지

# 01. Train Data 파악



가로 이미지



흐린 이미지



어두운 이미지

## 01. Train Data 파악

#	A	B	C	D	E	F	G	H	I	J
1		text_direction	img_direction	bbox_over	bbox_overlap	wrinkle	problem	bg	blur	stamp
2		세로 쓰기가 하나로 묶여 있으면 O 세로 쓰기가 따로 있으면 X	세로방향이면 vertical 가로방향이면 horizontal	text가 bbox를 넘어갔으면 O text가 bbox를 넘지 않았으면 X	겹쳐진 bbox가 있으면 O 겹쳐진 bbox가 없으면 X	종이가 구겨져 있으면 O 종이가 구겨져 있지 않으면 X	종이를 이미지로 옮기면서 생긴 문제가 있으면 O 종이를 이미지로 옮기면서 생긴 문제가 없으면 X	이미지에 bg가 있으면 O 이미지에 bg가 없으면 X	이미지가 흐리면 O 이미지가 흐리지 않으면 X	글씨들이 도장에 가려져 있으면 O 글씨들이 도장에 가려져 있지 않으면 X
3	drp.en_ko.in_house.deepnatural_003486.jpg	X	vertical	X	O	O	X	O	X	O
4	drp.en_ko.in_house.deepnatural_003489.jpg	O	vertical	X	X	X	X	X	X	O
5	drp.en_ko.in_house.deepnatural_003499.jpg	X	horizontal	X	X	X	X	O	X	O
6	drp.en_ko.in_house.deepnatural_003500.jpg	△	horizontal	X	X	X	X	O	X	O
7	drp.en_ko.in_house.deepnatural_003501.jpg	O	vertical	X	X	X	X	O	X	X
8	drp.en_ko.in_house.deepnatural_003502.jpg	X	vertical	X	O	O	X	O	X	O
9	drp.en_ko.in_house.deepnatural_003512.jpg	X	vertical	X	X	X	X	O	X	O
10	drp.en_ko.in_house.deepnatural_003516.jpg	O	vertical	X	X	O	O	X	X	O
11	drp.en_ko.in_house.deepnatural_003517.jpg	X	horizontal	X	X	X	X	X	X	O

#	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	stamp	text_cut	no_bbox	wrong_bbox	black_box	etc							
2	글씨들이 도장에 가려져 있으면 O 글씨들이 도장에 가려져 있지 않으면 X	잘린 글씨들이 bbox처리 되어 있으면 O 잘린 글씨들이 bbox처리 되어 있지 않으면 X	bbox가 없는 경우가 있으면 O bbox가 모두 있으면 X	bbox가 잘못 그려진 경우가 있으면 O bbox가 잘못 그려진 경우가 없으면 X	black box가 처리 되어 있으면 O black box가 처리되어 있지 않으면 X								
3	O	X	X	O	O								
4	O	X	X	X	X								
5	O	X	X	O	O								
6	X	X	X	O	X	다른 black box들을 bbox가 안그려져 있는데 하나의 bbox가 글씨들과 함께 하나의 bbox로 그려져 있음.							
7	X	X	X	O	O	도장 tag를 지우면 겹쳐진 bbox가 없음							
8	O	X	X	O	O								
9	O	X	X	X	X								
10	O	X	X	X	X								
11	O	X	X	O	O								



---

# 3. Data Relabeling

---

데이터 Relabeling 파이프라인

# 01. Annotation Guide

## Version \*\*\*

☰ Annotation

☰ Version 1.0.0

폐기	글자 없음
폐기	모든 글자가 알아보기 어려움
폐기	외국어가 전체 글자 영역의 1/2 이상
제외	번짐, 가려짐, 뭉개짐으로 글자 파악 불가
박스	모든 글자가 최소한의 박스에 들어gå야 함
박스	기본적으로 직사각형
박스	폰트 특성상 글자 크기 차이가 있다면 직사각형 변경 가능
박스	문장부호가 단독 주석된 경우 앞뒤 박스 높이 맞추기
박스 분리	어절 단위의 띄어쓰기
박스 분리	자간이 너무 넓거나 애마한 경우 : 글자 가로폭의 1/2 보다 넓으면 박스 분리
박스 분리	문장부호의 경우 바로 앞, 뒤 글자 폭 기준 1/2

## 02. Relabel Using CVAT(UFO ↔ COCO)

### UFO

```
"image_001.jpg": {
  "words": {
    "0": {
      "transcription": "E",
      "points": [[30,30],[134,30],[134,58],[30,58]],
      "orientation": "Horizontal",
      "language": "en",
      "tags": [],
      "confidence": 0.9900000095367432,
      "illegibility": false
    },
    "1": {
      "transcription": "gilldong.hong@upstage.ai",
      "points": [[30,93],[61,93],[61,112],[30,112]],
      "orientation": "Horizontal",
      "language": "en",
      "tags": [],
      "confidence": 0.9900000095367432,
      "illegibility": false
    }
  },
  ...
}
```

© NAVER Connect Foundation

### COCO

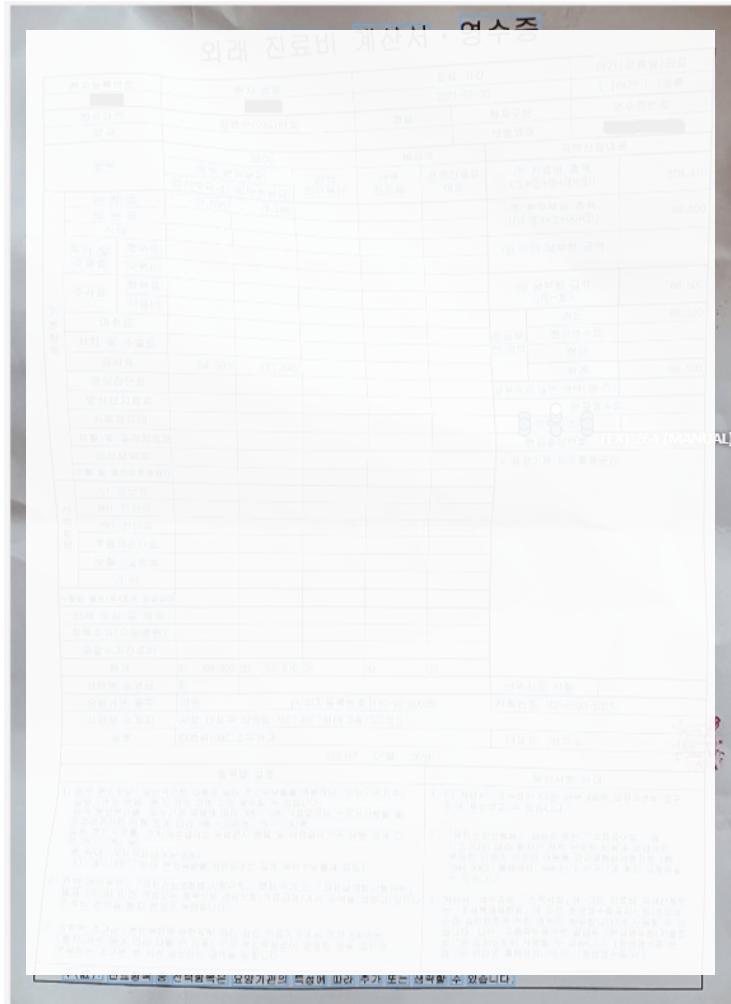
```
image{
  "id"           : int,
  "width"        : int,
  "height"       : int,
  "file_name"    : str,
  "license"      : int,
  "flickr_url"   : str,
  "coco_url"     : str,
  "date_captured": datetime,
}
```

```
annotation{
  "id"           : int,
  "image_id"     : int,
  "category_id"  : int,
  "segmentation" : RLE or [polygon],
  "area"         : float,
  "bbox"         : [x,y,width,height],
  "iscrowd"      : 0 or 1,
}

categories[{
  "id"           : int,
  "name"         : str,
  "supercategory": str,
}]
```

- Categories와 기타 메타정보는 placeholder로 대체
- 포맷에 맞게 좌표 변환
- points(UFO) -> bbox(COCO)로 변환했으나,  
polygon의 vertices를 이용하는 segmentation으로 변환시키는 게 더 쉬울 것으로 예상 됨

## 03. Relabel Using CVAT



- 앞에서 정한 annotation 가이드를 숙  
지하고 relabeling 진행
- 팀원 5명에게 60장씩 분배

## 04. Relabel Using CVAT (Workspace)





Organization: AITech03

GEN잡아 [🔗](#)  
OCR 리라벨링 [🔗](#)

📞 Add phone number [🔗](#)  
✉ Add email [🔗](#)  
📍 Add location [🔗](#)  
Created May 28th 2023  
Updated 4 days ago

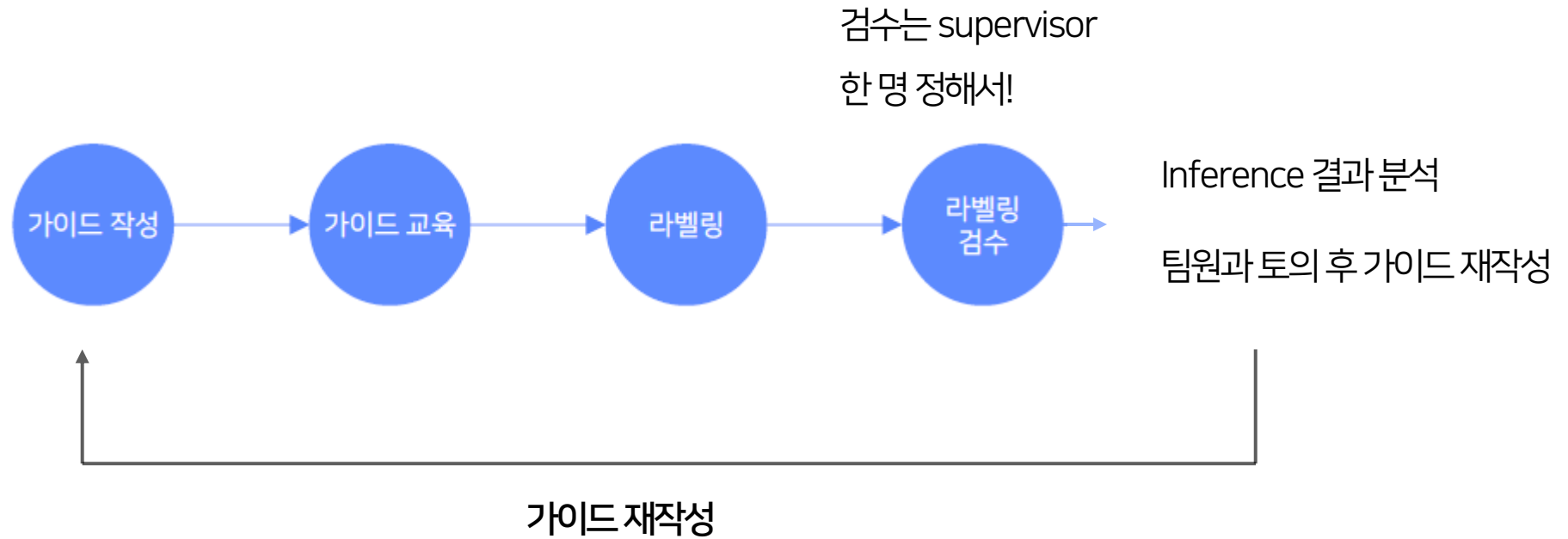
Actions ⋮

👤 Invite members

		Invited 3 days ago by seungki Joined 3 days ago	Worker ▼	🗑
		Invited 3 days ago by seungki Joined 3 days ago	Worker ▼	🗑
		Joined 4 days ago	Owner ▼	

- 하나의 작업 위에서 동시에 라벨링 해야 할 경우  
Organization에서 팀원 초대 (무료는 3명까지)
- 프로젝트를 진행 할 때는 개인 workspace에서 진행한  
경우가 많음!

## 05. Data Relabeling 파이프라인



- Inference에서 특수 case, 잘못 예측하는 부분에 대한 분석
- 주의 깊게 볼 케이스에 대한 가이드라인 재작성
- 일관되지 않은 라벨링이 보일 경우 가이드 교육 후 라벨링 작업 재분배

---

# 4. 실험

---

실험 시간 단축  
데이터 Re-labeling  
실험 내용

## 01. 실험 시간 단축

```
class SceneTextDataset(Dataset):
    def __init__(
        self,
        root_dir,
        split="train",
        num=3,
        image_size=1024,
        crop_size=512,
        color_jitter=True,
        normalize=True,
    ):
        if num == 0:
            pkl_dir = osp.join(root_dir, "ufo/{}.pickle".format(split))
        else:
            pkl_dir = osp.join(root_dir, "ufo/{}.pickle".format(split + str(num)))

        with open(pkl_dir, "rb") as fr:
            total = pickle.load(fr)
```

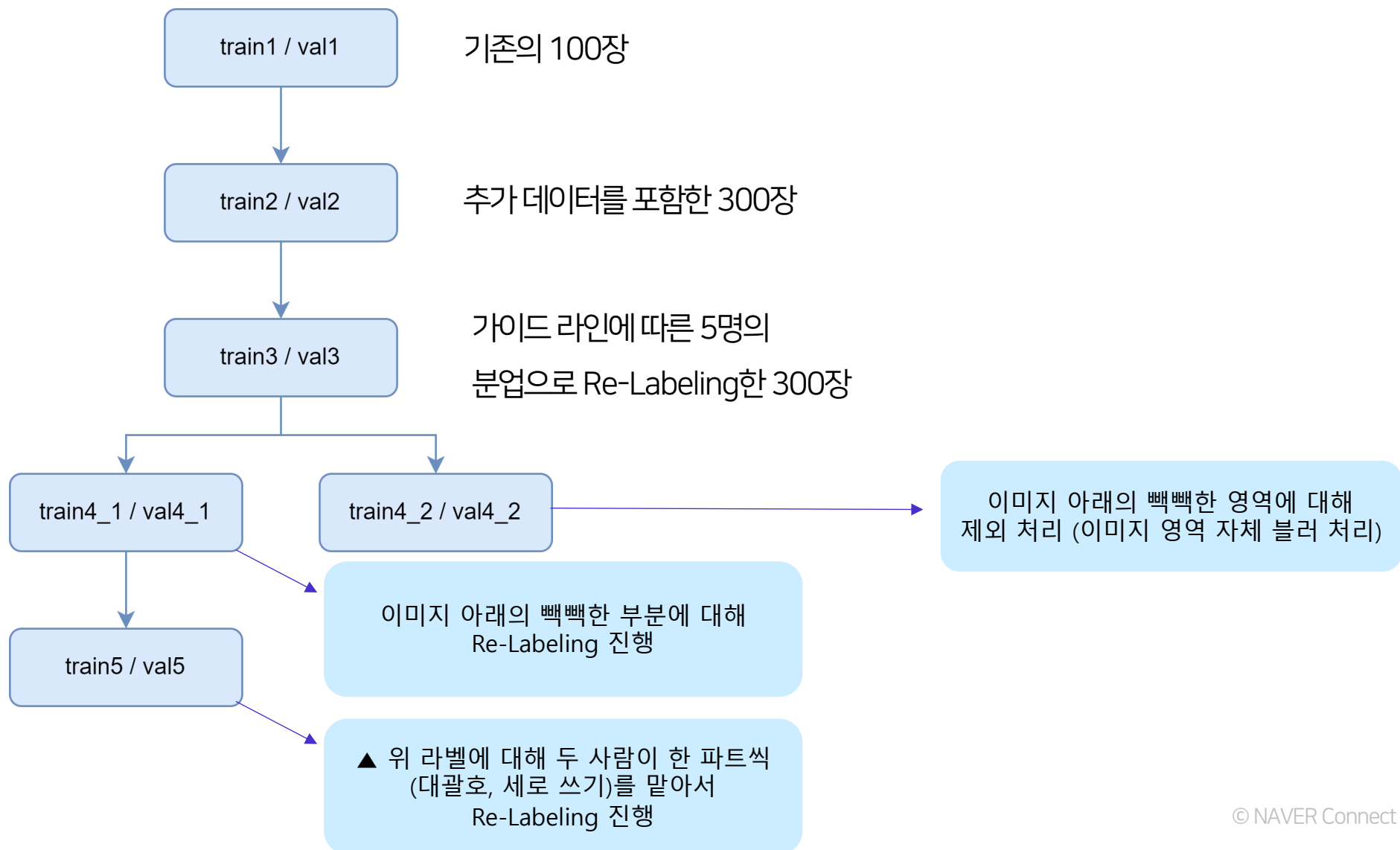
13~14분



6~7분



## 02. Data Relabeling



## 03. 실험 내용

베이스라인 코드  
실험

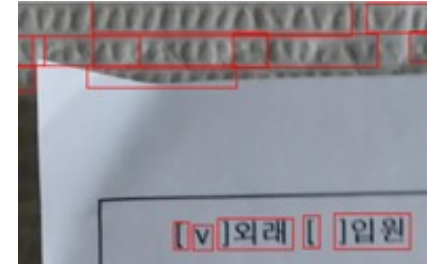


Inference 결과 파악



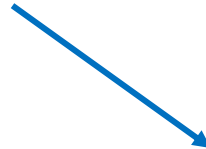
배경을 검출하는 경우

- 배경 이미지만 있는 데이터셋을 이용해 Fine-Tuning



QR 코드 검출하는 경우

- 항목별 설명, 일반 사항 안내 Blur
- 해당하는 annotation 전부 삭제



세로 쓰여진 가로쓰기 검출 못하는 경우

- ?



seed : 1004  
epoch : 150  
f1 score : 0.9546

---

# 5. Ensemble

---

앙상블

## 01. WBF 앙상블

기존 WBF  
F1 score  
0.9664 -> 0.9498

- 한 모델에서 잘못 검출하는 경우가 있어도 다른 한 모델이 검출하지 않으면 결과에서 제거
- 앙상블 과정에서 bbox가 이전보다 더 정교해짐

## 01. WBF 앙상블

수정 후 WBF  
F1 score  
0.9664 -> 0.9848



+



=



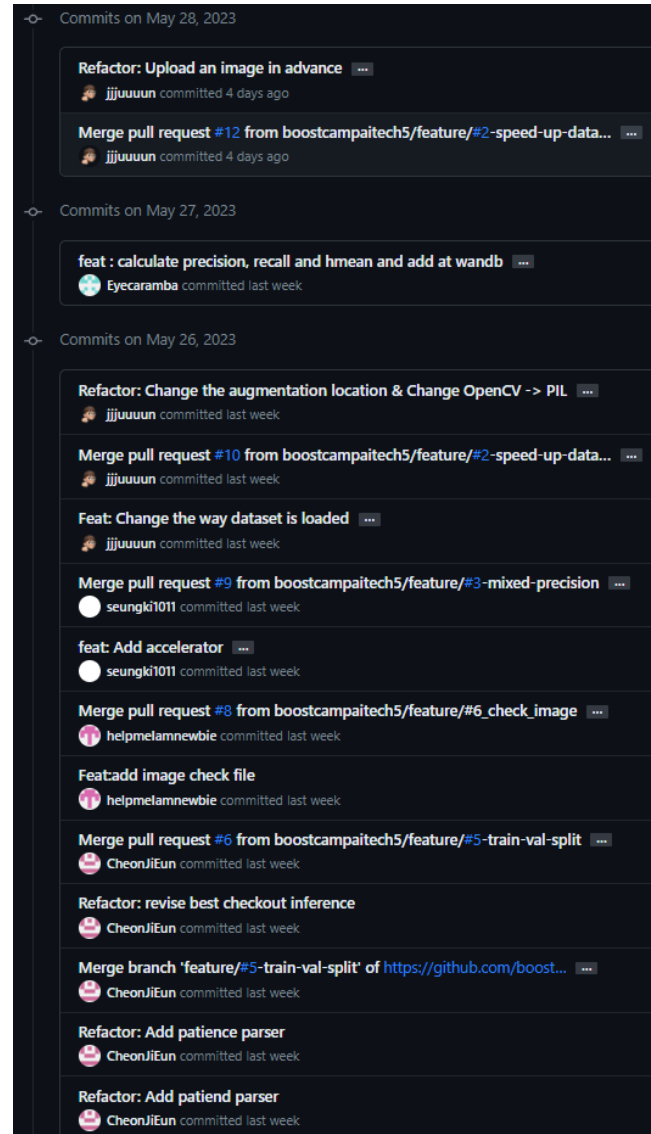
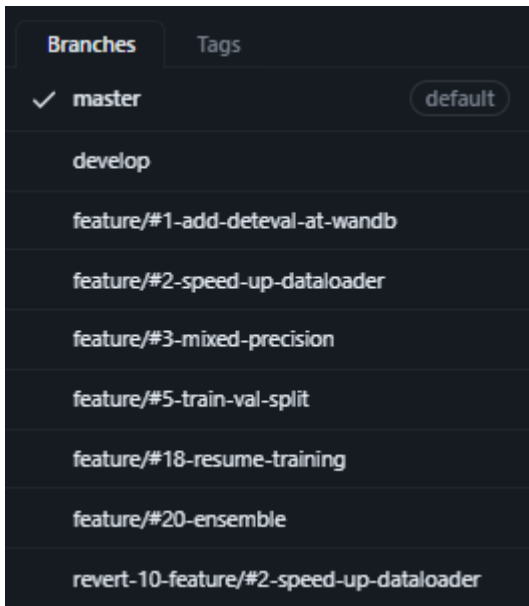
---

# 6. 협업 방식

---


협업 방식 소개

# 01. Git Convention




- Git Flow를 이용한 브랜치 전략
- Udacity convention 기반의 commit convention


## 02. Notion


 [오피스 아워 정리 \(EAST 모델\)](#)


 [detect.py 코드분석](#)


 [JSON 파일 공유](#)

 [CVAT 사용법](#)

 [INFERENCE 분석](#)

 [Re-labeling](#)


 [COCO JSON](#)

 [외부 데이터셋](#)


 [Code](#)


 [ENSEMBLE](#)

 [EDA](#)


 [OCR 대회 요약](#)

 [GitHub](#)

 [기존 연구](#)

 [가이드 라인](#)

 [Augmentation](#)

 [Wrap-up report](#)

- 목표
  - Data-Centric 대회 답게 데이터를 최대한 만져보자.
- 추가 작업
  - 추가할 데이터 셋 찾기
- 기본 셋팅 값
  - Optimizer
    - 종류 → Adam

- Notion에서 각자 실험, 작업한 내용 공유



---

End of Document  
Thank You.