

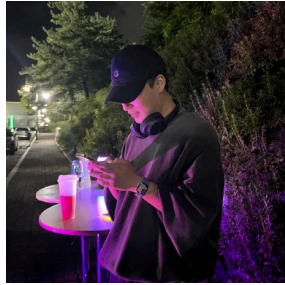
Document Type Classification | 문서 타입 분류 대회

12조



김나리

팀장, 발표, EDA,
Pre-processing,
Data
Augmentation,
Modeling



박범철

EDA,
Modeling,
OCR



서혜교

EDA, Pre-
processing, Data
Argumentation,
Modeling



조용중

EDA, Pre-
processing, Data
Augmentation,
Modeling, OCR



최윤설

EDA, Pre-
processing,
Data
Augmentation,
Modeling

Competiton Info

Overview

이번 대회는 computer vision domain에서 가장 중요한 태스크인 이미지 분류 대회입니다.

이번 대회를 통해서 문서 타입 데이터셋을 이용해 이미지 분류를 모델을 구축합니다. 주어진 문서 이미지를 입력 받아 17개의 클래스 중 정답을 예측하게 됩니다. computer vision에서 중요한 backbone 모델들을 실제 활용해보고, 좋은 성능을 가지는 모델을 개발할 수 있습니다. 그 밖에 학습했던 여러 테크닉들을 적용해 볼 수 있습니다.

Timeline

- 2024년 7월 29일 : 대회시작 각자 데이터 EDA
- 2024년 7월 30일 ~ 8월 2일 : 온라인 강의, 데이터 Augmentation을 이용한 모델링, Baseline code 학습
- 2024년 8월 5일 : Swin Transform, Convnext v2 적용
- 2024년 8월 6일 : OCR, Augrpy 코드 공유 적용
- 2024년 8월 7일 : 데이터 오프라인 증강, Test data의 Denoising 적용
- 2024년 8월 8일 : 각자의 모델 Hyper parameter tuning, LM3 적용
- 2024년 8월 9일 ~ 11일 : 각자의 모델 학습시키면서 리더보드 올리기

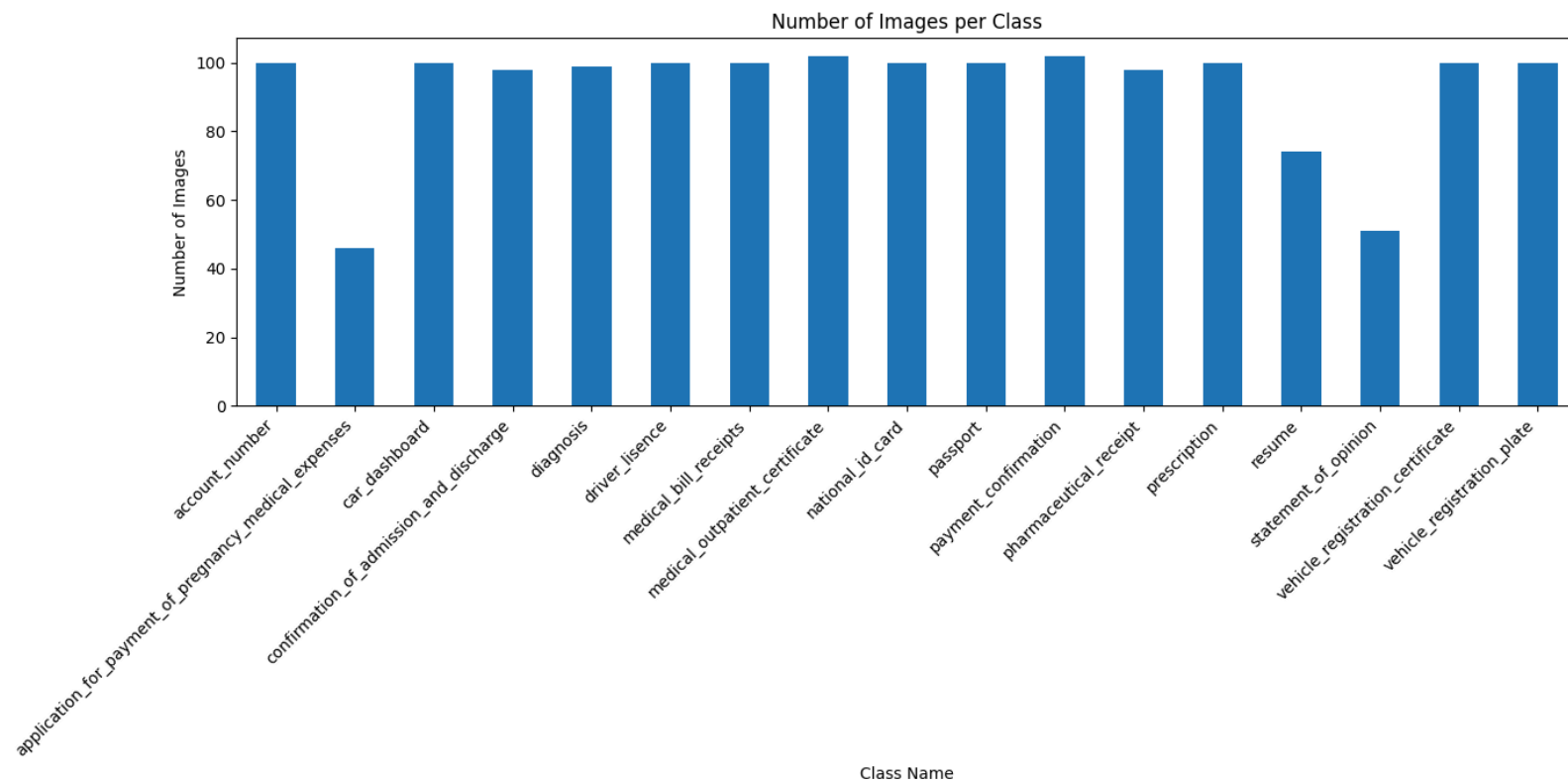
Data description

Dataset overview

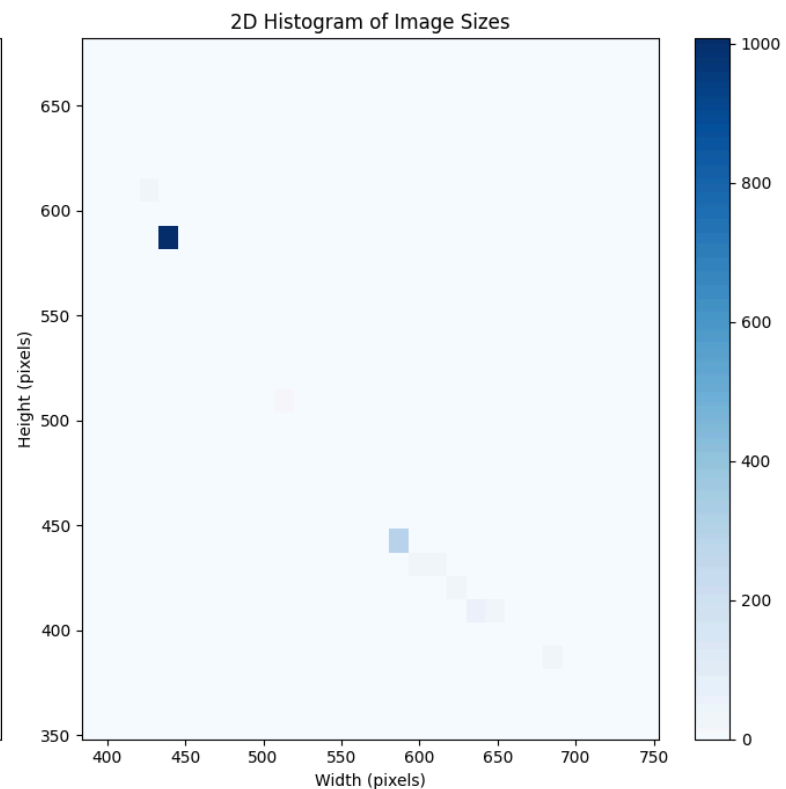
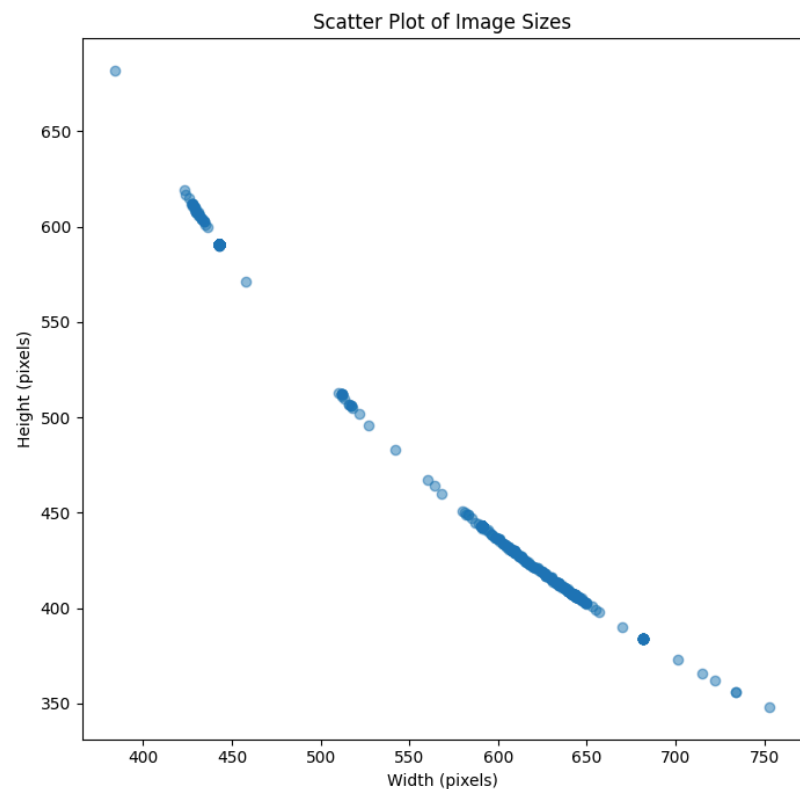
- train [폴더]: 1570장의 이미지가 저장.
- test [폴더]: 3140장의 이미지가 저장.
- train.csv [파일]: 1570개의 행으로 이루어져 있습니다. train 폴더에 존재하는 1570개의 이미지에 대한 정답 클래스를 제공.
 - ID: 학습 샘플의 파일명
 - target: 학습 샘플의 정답 클래스 번호
- sample_submission.csv [파일]: test 폴더에 존재하는 3140장의 이미지에 대한 제출샘플 제공. train.csv와 같은 형식으로 이루어져있음.
- meta.csv : 17개 클래스에 대한 설명

EDA

- 학습 데이터는 대체로 clean한 반면 평가 데이터는 상/하/좌/우 반전 및 회전등이 적용된 noise 데이터
- class 별 학습데이터의 양이 고르지 못함.



- 직사각형 이미지 99%
- 이미지들의 사이즈 분포 시각화.



Data Processing

- 학습 데이터에 오분류된 데이터 확인하여 label 수정

38, 약제비 영수증

192, 소견서

Train Dataset Incorrect Label

340, 약제비 영수증

428, 입퇴원 확인서

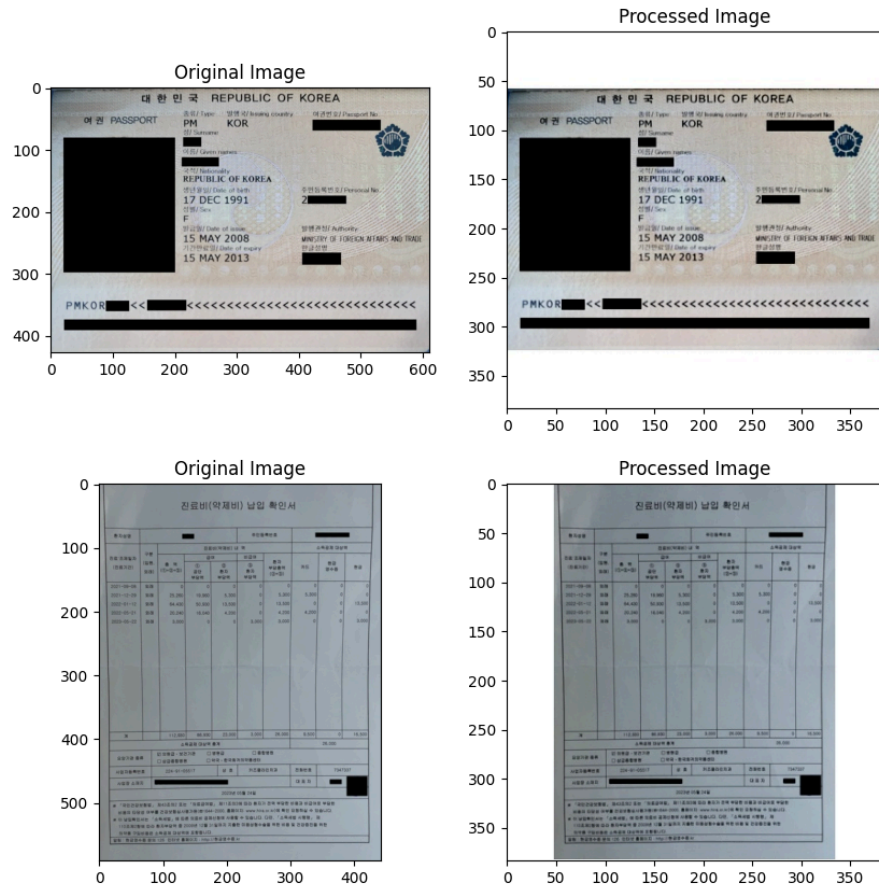
723, 입퇴원 확인서

862, 통원/진료 확인서

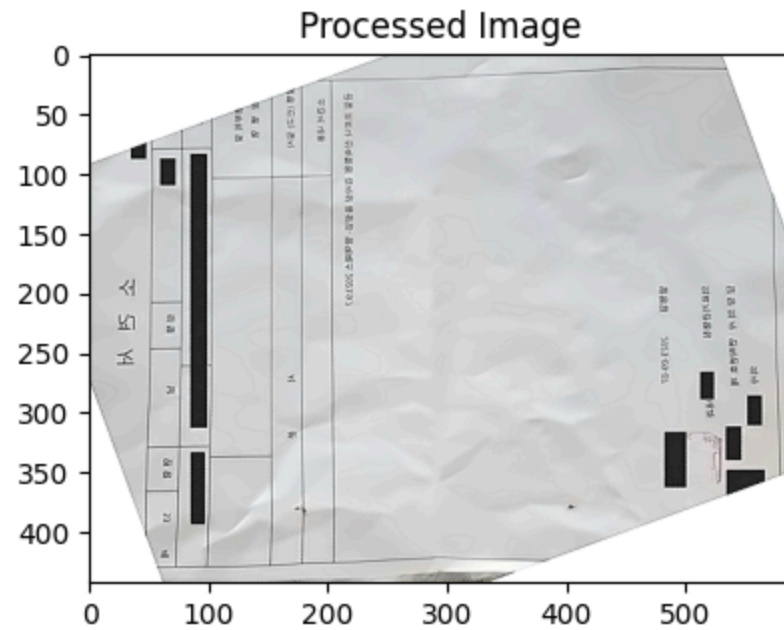
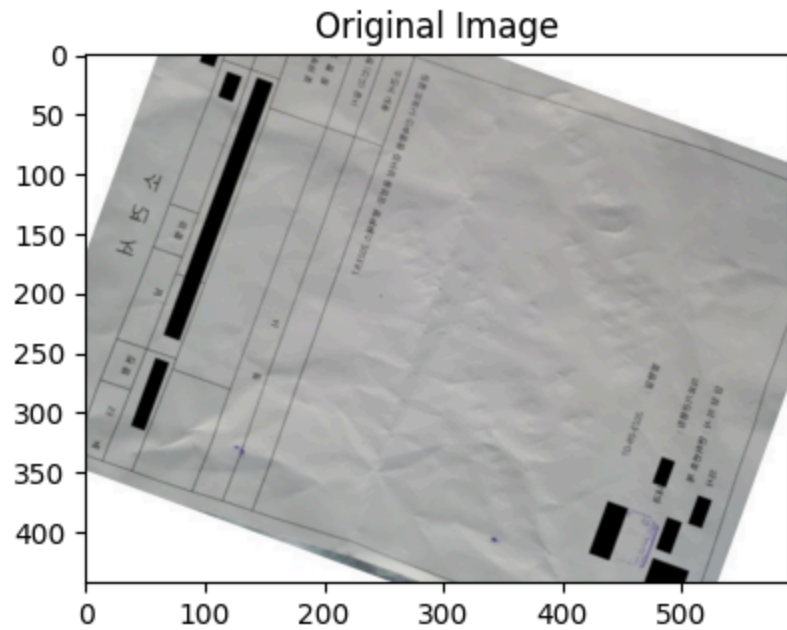
1095, 입퇴원 확인서

1237, 진단서

- 이미지의 크기를 모델의 최적사이즈에 맞추기 좋게 이미지를 가운데에 두고 여백을 주면서 정방형으로 만든 다음 리사이즈.



- 이미지의 회전 바로잡기(Denoising)



Data Augmentation

- 평가 데이터 셋에 대한 분석을 통해 'augraphy' 의 다양한 기능 적용
 - i. 윤곽선 감지를 사용하여 텍스트 선을 감지하고 부드러운 텍스트 취소선, 강조 또는 밑줄 효과 추가
 - ii. 이미지에 낙서 적용
 - iii. 입력 용지의 색상 변경
 - iv. 잉크 번짐 효과 (두 이미지 혼합하여 블리드스루 효과)
 - v. 접기 효과
 - vi. 조명 또는 밝기 그래디언트
 - vii. 종이 표면에 그림자 효과
 - viii. 크기 조정(resizing), 뒤집기(flips), 회전(rotation) 등 기본적인 기하학적 변환 적용
- torchvision.transforms v1과 호환되는 v2 사용

- 적용 후

After Data Augmentation

Label: 5



Label: 9



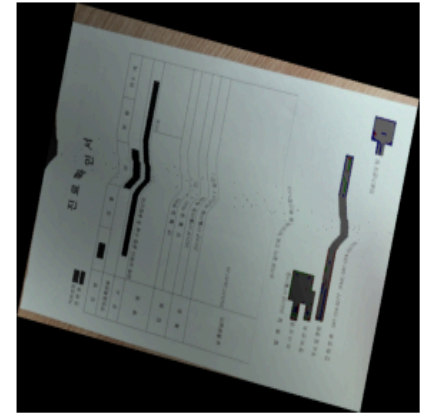
Label: 2



Label: 2



Label: 7



Modeling

Model description

- EfficientNet_b4, V2
- SWIN(Shifted Window) Transformer
- ConvNeXt V2
- Paddle OCR

Modeling Process

1. Augmentation으로 데이터증강하여 3개의 모델 실험
2. 하이퍼파라미터 튜닝
3. 데이터를 오프라인으로 증강시켜 학습. (약 25000개)
4. 평가데이터 Denoising
5. 훈련데이터중 일부도 Denosing
6. Paddle OCR을 이용한 단어 추출 후 단어사전을 만들어 분류 (3, 4, 7, 14 클래스만 적용함)

<https://api.wandb.ai/links/narykkim/p2l1gyy0>


Result

Leader Board

- 리더보드[중간 순위]

순위	팀이름	팀멤버	F1 score	제출 횟수	최종 제 출
내등수 4	cv12		0.9530	85	16h

- 리더보드[최종 순위]

순위	팀이름	팀멤버	F1 score	제출 횟수	최종 제 출
내등수 4	cv12		0.9361	85	16h

후기

- 김나리 : 데이터를 이미지와 문서로 분류해서 다시 분류하는 시스템을 만들고 싶었는데, 초반에 데이터량이 적다보니 좋은 결과가 나오지 않아 중단했다. 이미지를 오프라인으로 증강한 후에 시도했으면 좋았을텐데 그러지 못해 아쉽다. Paddle OCR을 시간관계상 깊이 공부하지 못하고 지나갔는데, 좀더 체계적으로 만들어보고싶다. OCR 대회를 기대해본다.
- 박범철 : Confusion Matrix에서 단일 분류모델에서 FN, FP를 가져와 OCR을 도전했지만, 출력 크기 오류 때문에 시간을 많이 잡아먹었던 점이 아쉽고, OCR 관련하여 오류를 고치고 계속 시도해봤다면, 세 가지 모델을 앙상블 하였을 때 배깅 비율을 다르게 했었으면. 데이터 증강쪽을 좀 더 전문적으로 실시했다면 에 대한 아쉬움.
- 서혜교 : Augraphy 제대로 구현해보기, SwinT 논문부터 제대로 심도깊게 읽고 리뷰하기.

- 조용중 : OCR 부분을 시도했지만 성공적인 결과는 못 얻은점, model 부분에서 Pretrained=True/False 에 대해서 충분히 테스트를 하지 못한점이 아쉽고, 최신 모델을 논문 참조하여 로우 레벨로 구현해 보는것. 오류가 큰 하위 몇개 클래스들에 대해서 계층적으로 가중치를 주고 모델에 적용하는 것을 시도해보고 싶다. 초기 Preprocessing 을 좀더 다양하게 시도하자. test 데이터를 꼼꼼히 살펴볼 것.
- 최윤설 : 하다보니 이것저것 시도해보고 싶은게 많았는데 시간이 부족해서 아쉬웠음, 다음 대회때 부터는 대회 오픈하자마자 이것저것 해보기, 다양한 라이브러리를 통해 이미지에 noise를 추가하여 실 데이터와 같이 변형시킬 수 있음. 매 대회를 진행하면서 느끼는 점은 데이터 전처리의 중요성! Pytorch lightning + hydra 로의 변환을 시도 해보고 싶다.

인사이트

- 김나리 : Augraphy나 layoutLM 등 조원님들이 공부해서 공유해주신 소중한 샘플코드
- 박범철 : 세 가지 모델을 앙상블 하였을 때 배깅 비율
- 서혜교 : Augraphy 제대로 구현해보기, SwinT 논문부터 제대로 읽고 코드작성하기
- 조용중 : 초기 Preprocessing 을 좀더 다양하게 시도했었으면. test 데이터를 꼼꼼히 살펴볼것.
- 최윤설 : 다양한 라이브러리를 통해 이미지에 noise를 추가하여 실 데이터와 같이 변형시킬 수 있음. 매 대회를 진행하면서 느끼는 점은 데이터 전처리의 중요성!

시도해 보고 싶은 점

- 김나리 : Paddle OCR을 시간관계상 대충하고 지나갔는 데, 좀더 체계적으로 만들어보고싶다. OCR 대회를 기대해본다.
- 박범철 : OCR 관련하여 오류를 고치고 계속 시도해봤다면, 세 가지 모델을 앙상블 하였을 때 배깅 비율을 다르게 했었으면. 데이터 증강쪽을 좀 더 전문적으로 실시했다면.
- 서혜교 : 논문을 심도깊게 읽고 리뷰하기.
- 조용중 : 최신 모델을 논문 참조하여 로우 레벨로 구현해 보는것. 오류가 큰 하위 몇개 클래스들에 대해서 계층적으로 가중치를 주고 모델에 적용해 보는것.
- 최윤설 : Pytorch lightning + hydra 로의 변환