OCR을 이용한 문장 인식 모델 구현

Team: 권정을, 이시내, 윤선영, 이찬, 정재원

BOAZ

01 프로젝트 주제

✓ 기존 프로젝트 주제 : 한국어 책 제목 인식을 위한 STR 구현

✓ 변경된 프로젝트 주제 : OCR을 이용한 문장 인식 모델 구현

✓ 변경 이유

① STR은 자연환경에서 발생하는 텍스트 인식에 중점을 두게 되는데, 이러한 데이터의 양이 많이 없음

② STR의 모델의 사이즈가 커서 로컬에서 학습하기가 어려움이 있음

③ 한국어로 학습된 모델이 없어서 학습시키기에 까다로움

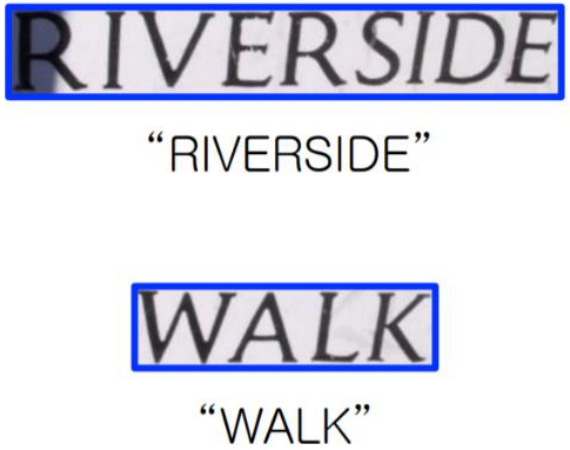
→ OCR task를 수행할 수 있는 한국어 모델 다수 존재, OCR을 위한 전처리가 되어 있는 공공 데이터 습득 가능

01 프로젝트 주제

✓ OCR이란 ?

이미지 형태를 읽어서 데이터의 내용을 분석하고 그림 영역과 글자 영역으로 구분한 후, 글자 영역의 문자들을 텍스트의 형태로 변환하여 주는 것

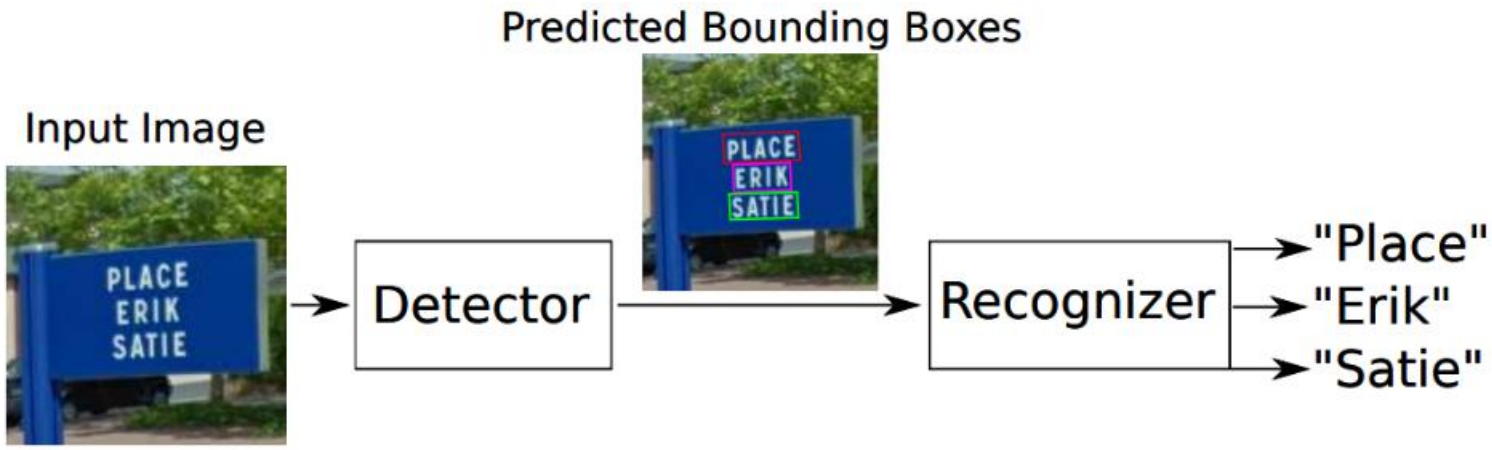
Ex) 스마트폰으로 카드결제를 진행할 때, 카메라로 카드를 인식하면 자동으로 카드 번호가 입력되는 것



02 프로젝트 소개

"OCR = Text detection + Text recognition"

딥러닝 기반의 OCR은 아래 그림에서 보이는 것처럼, 크게 이미지 속 문자를 읽는 문자 영역 검출 Text detection과 검출된 영역의 문자를 인식하는 Text Recognition으로 구분할 수 있음



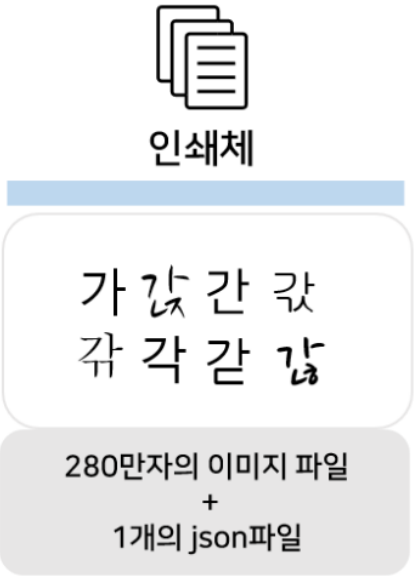
03 실험

✓ OCR을 위해 제안된 두 가지 모델의 성능을 비교

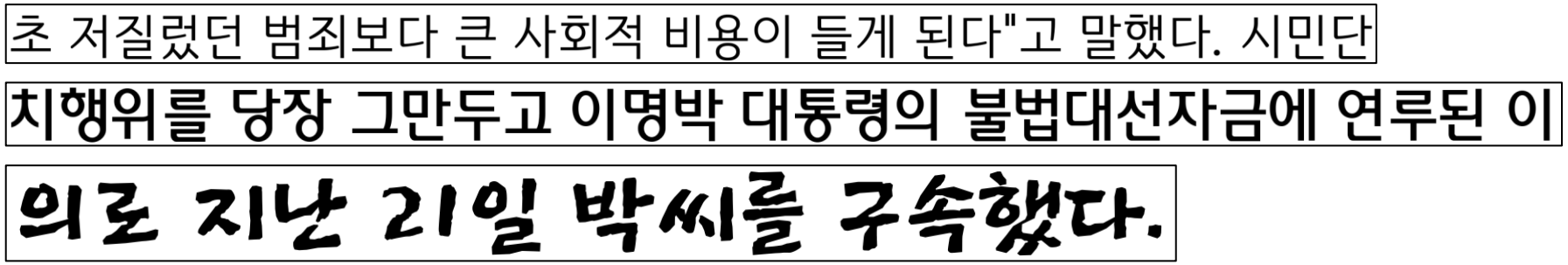
1. TrOCR
2. Pororo

✓ 사용 데이터셋

- Ai hub - 한국어 글자체 이미지 사용
- 인쇄체 이미지 데이터(문장)



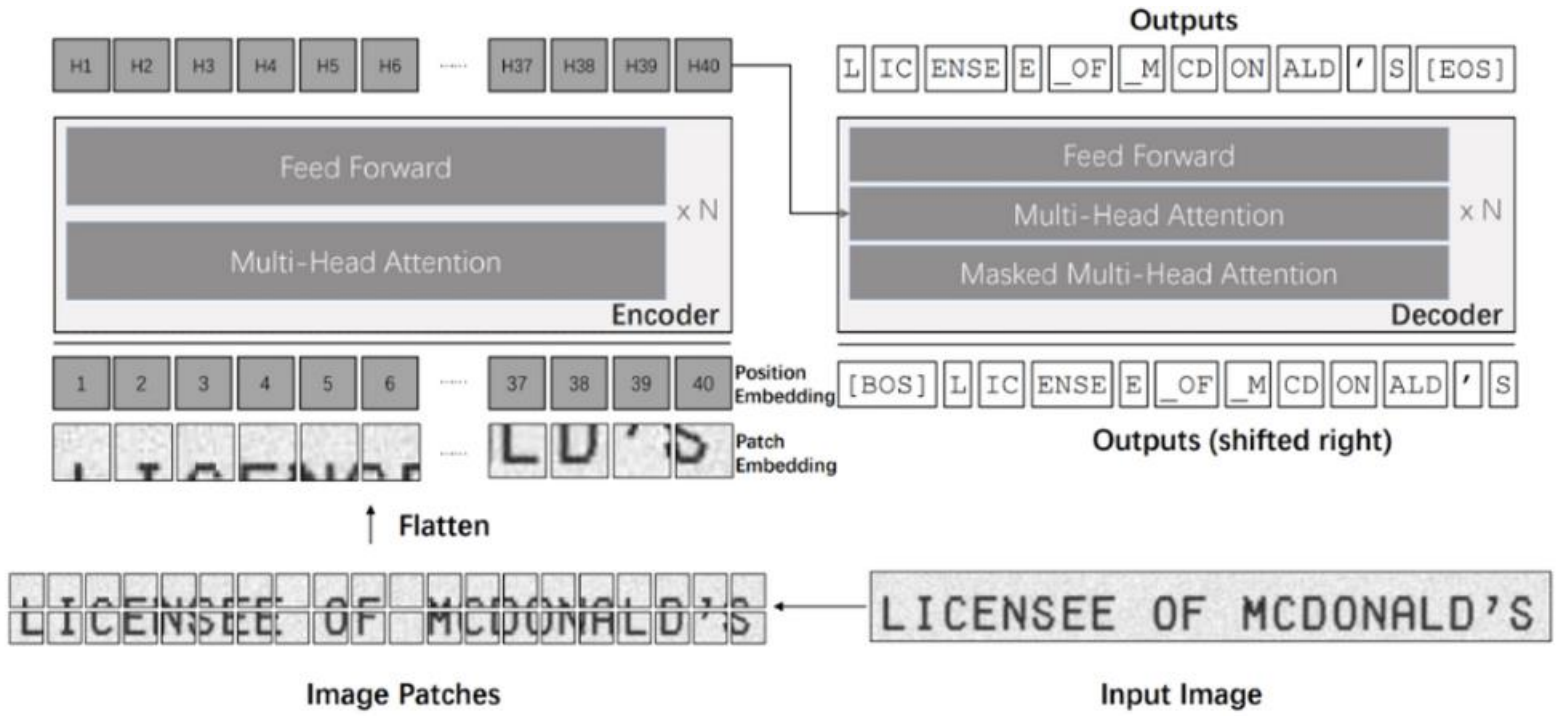
[데이터 예시]



03 실험

TrOCR (Transformer-based Optical Character Recognition)

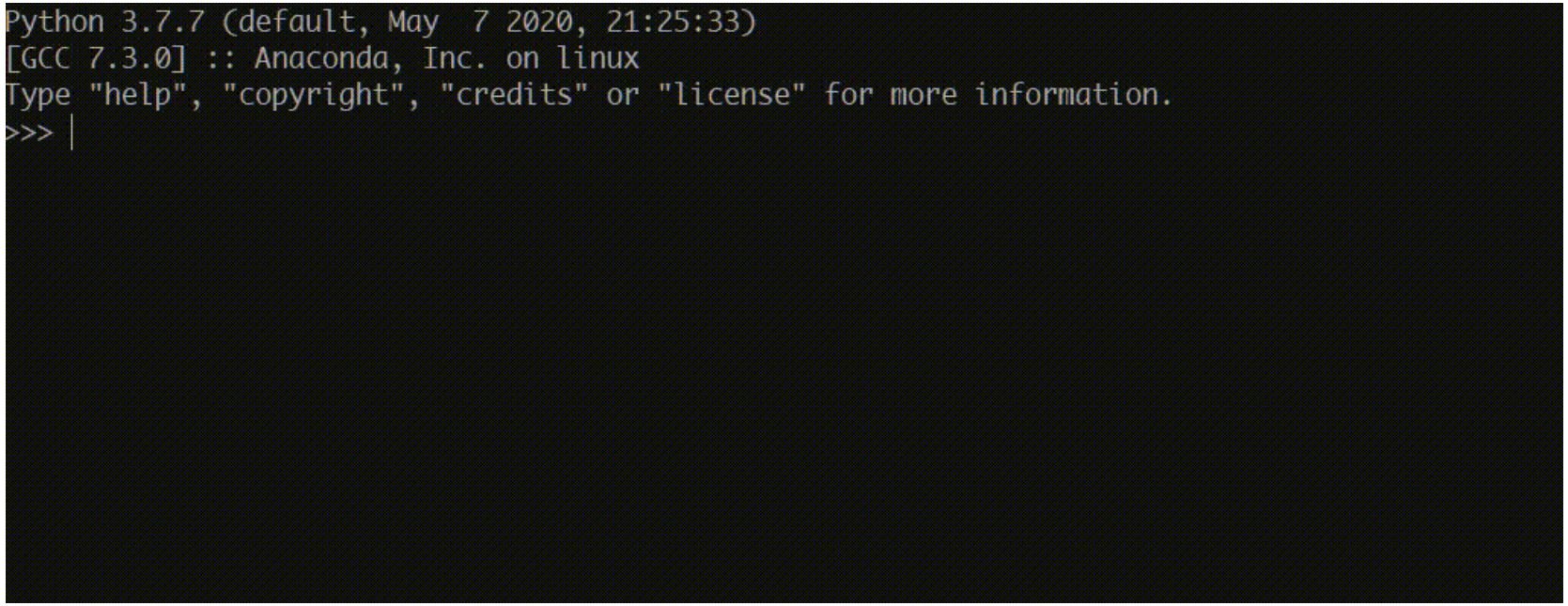
- ✓ 기존 모델의 복잡한 전/후처리 과정을 개선한 Transformer 기반의 end-to-end text recognition 모델
- ✓ TrOCR에서는 입력 이미지를 정해진 크기(patch)로 나눈 후 patch sequence를 만들어서 transformer에 적용해준다



03 실험

Pororo OCR(Platform Of neuRal mOdels for natuRal language prOcessing)

- ✓ Pororo는 카카오에서 카카오브레인에서 공개한 오픈소스 라이브러리이다
- ✓ 다양한 기능을 제공하지만 그 중 필요한 OCR 기능만을 추출하여 간단하게 사용할 수 있도록 만들었다



04 결과

	outputs	label
0	불만과 불쾌감을 공개적으로 언급해 남북관계를 더욱 경색시키고 있	불만과 불쾌감을 공개적으로 언급해 남북관계를 더욱 경색시키고 있
1	대 의예과 1515학번 남학생 11명은 지난해 35월 학교 인근 그집집	대 의예과 1516학번 남학생 11명은 지난해 35월 학교 인근 고깃집
2	시 김종창 금감원장에게 청탁한 정황을 포착한 것으로 30일 확인됐	시 김종창 금감원장에게 청탁한 정황을 포착한 것으로 30일 확인됐
3	으로 알려지고 있다. 지난 5월 8일 구미시와 국가공단의 생명줄인 해	으로 알려지고 있다. 지난 5월 8일 구미시와 국가공단의 생명줄인 해
4	고의 스파클링 와인이다. 대표적인 스테미너 음식으로 꼽히는 장바구니	고의 스파클링 와인이다. 대표적인 스테미너 음식으로 꼽히는 장어구
...	...	...
840	각 중단하고 박 원내대표에게 혐의가 있다면 당당히 기소하라 " 며 "	각 중단하고 박 원내대표에게 혐의가 있다면 당당히 기소하라"며 "한
841	등 번개가 치겠습니다. 비는 밤에 서쪽 지역부터 차차 그치겠습니.	등 번개가 치겠습니다. 비는 밤에 서쪽 지역부터 차차 그치겠습니다.
842	골인하게 됐다. 예비 신랑은 1살 연상의 사업가 장모씨씨로 고대	골인하게 됐다. 예비 신랑은 1살 연상의 사업가 장모(32)씨로 고대
843	된다. 고무에 카본블랙주1을 섞은 일반타이어는 날씨가 추워질수록	된다. 고무에 카본블랙주1을 섞은 일반타이어는 날씨가 추워질수록
844	법규위에 계류돼 있는 법안에 북한에 대한 인도적인 지원 내용이 들	법사위에 계류돼 있는 법안에 북한에 대한 인도적인 지원 내용이 들

845 rows × 2 columns

잘 된 예시)

Input image **불만과 불쾌감을 공개적으로 언급해 남북관계를 더욱 경색시키고 있**

Inference text “불만과 불쾌감을 공개적으로 언급해 남북관계를 더욱 경색시키고 있”

잘 안된 예시)

Input image 리도 있다.

Inference text “라고 있다.”

04 결과

성능 평가: BLEU Score(Bilingual Evaluation Understudy Score)

- ✓ BLEU는 기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지 비교하여 번역에 대한 성능을 측정하는 방법
- ✓ OCR task가 잘 수행되었는지 판단하기 위해 GT(Ground Truth)와 모델 Output을 바탕으로 BLEU score를 계산
- ✓ Aihub 데이터 중 845개의 샘플을 Test set으로 설정하여 이에 대한 OCR performance를 측정함

Model	BLEU score
TrOCR	0.75
Pororo	**0.81**

Transformer 기반의 TrOCR 보다,
CNN(CRAFT) 기반의 Pororo가
image 내 text의 특징을 더 잘 학습할 수 있기 때문에 성능이 더 높았다고 생각됨

BOAZ

05 결론 및 한계점

- ✓ 사용한 데이터의 경우 OCR의 진정한 목표인 자연 상태에서의 글자 인식이 아닌 정제된 형태로 OCR의 성능 평가로 적절하지 않았다
- ✓ bleu score 외의 다른 측정 matric을 사용해보지 못한 점이 아쉽다
- ✓ pororo의 경우 fine tuning등의 모델에 직접적인 접근을 해보지 못해 사용한 데이터에 fit하지 못한점에서 bleu score가 아쉽게 나왔다

감사합니다