2022.09.15(목) 3주차 진행

안팀장님

1) 만들어진 학습모델로 테스트 진행

2) 다른 더 좋은 모델로 갈아껴서 학습진행

3) 밑줄, 볼드체 등의 데이터 생성방법 찾아보기

4) 타겟 데이터 수집 (데이터 에듀 영상에서 따오기)

5) ANN모델 + 전처리모델 연결 (완료)

인식모델에 학습에 필요한 데이터 생성 오픈소스 프로젝트

TextRecognitionDataGenerator

위 프로젝트에서 기본적으로 제공해주는 한글소스 10%

자주 쓰이는 한국어 낱말 5800 20%, 조합해서 낱말 두개 25%

국민대 교수가 온라인 배포하는 한국어 현대문구 모음집에서

문장을 어절단위로 끊어서 한어절만듬 20%, 조합해서 두어절씩 25%

데이터 총 100만건, train 75%, val 25%

easy OCR이 한글을 인식할때 사용하는 모델구조 : None-VGG-BiLSTM-Seed

현재 easy OCR에서 한글을 인식하는 모델 : korean\_g2

OCR 학습툴을 제공하는 오픈소스 프로젝트

deep-text-recognition-benchmark

위 프로젝트에서 best accuracy라고 하는 모델구조 :TPS-ResNet-BiLSTM-Attn

근데 TPS-ResNet-BiLSTM-Attn 이걸 사용하려고 하니까

korean\_g2 얘의 구조가 None-VGG-BiLSTM-Seed라서 학습이 안됨

그래서 미세조정(fine tuning)을 하는 방향으로 가려면 korean\_g2모델에

None-VGG-BiLSTM-Seed 구조를 유지하면서 학습을 해야함

easy OCR에서 제공하는 trainer로 학습 진행중 (현재 진행중)

허사원

1) 모델업데이트1 : 특수문자 제거 함수 추가 / custom모델 연결 코드 추가 / ocr 문단 구분(완료)

2) 페이지별 고정 문단 내용\_배경자막 제거(진행중)

9/19 주차에 할 일

3) 이미지 유사도 검출 : 새로운 중복 제거 방법 함수 추가 \*\*

4) (문제점 발생 : korea\_g2모델 RGB 요소가 단일로 작동하여 흑백처리 필요) 그레이스케일 추가하기

5) 영어 맞춤법 검사기 : 한글 돌리고 영어 돌렷을때 영어검사기가 한글을 안 헤칠까? - 고민할 부분

6) 유튜브 타임스탬프와 형식 동일시하기

작은이사원

1) 두런 사이트 참고해서 음성 데이터 검색 방법(현재 구현중)

→ 텍스트데이터로 구현할 수 있는지 확인 (검색기능)

2) 텍스트 데이터 中 표, 그림 데이터 중간에 나타났을 경우 진행방향 확인

3) 현황분석 자료 확인

두런 사이트 소스 확인 = Youtube 사이트 연결방법 확인

→ 멘토님 질의사항으로 문의드리기(차이점 및 구현 방향)

Youtube URL:<https://youtu.be/g4ctGPOZV1M?list=PLno8wKm9WlnmYkbJkFYAOgt9> gDhB99Evf

DolearnURL:https://www.youtube.com/embed/g4ctGPOZV1M?enablejsapi=1&amp;origin=https%3A%2F%2Fwww.dolearn.ai&amp;widgetid=1

큰이사원

1) 플라스크를 통한 웹 사이트 구축

2) 전처리모델(함수) 플라스크 구현 프로세스 설계

1. 비디오 -> 프레임 -> 한글인식+타임스탬프 함수 (산출물 : 딕셔너리)
2. 딕셔너리 받아서 한국어 문법 교정 (산출물 : 리스트)
3. 리스트 중복제거 함수 (산출물 : 리스트)
4. 최종결과물 -> 리스트
5. 위 내용을 클래스화

김사원

1) easyocr과 tesseract 비교 정리 ppt 올리기

2) 큰 이사원과 콜라보

<프로젝트 일정>

09:10 - 09:30 회의

09:30 - 12:50 프로젝트 진행

14:00 - 14:50 프로젝트 공유 및 점검

15:00 - 15:50 프로젝트 진행

16:00 - 17:50 개인 공부