[KB국민은행] AI부분 전문직무직원 발표자료

윤여훈

주제 : 3) OCR 과제 (10분)

1. 시작
2. 3) OCR 과제 선정 이유(~45초)

주제 : (3번째 과제인) 이미지 파일에서 체크박스를 찾고 체크 유무를 판단할 수 있는 처리 방법 제시 및 구현

선정이유

* 이미지 처리 모델 프로세스 경험 및 기술 공유
* 프로젝트의 관련된 데이터 처리/분석 경험과 기술 공유
* 객체 검출 및 픽처 엔지니어링에 대한 과정과 파이프라인에 대한 기술 공유

1. 주제 / 목적 (1) ( ~ 2분 20초)

과제 목적

* 이미지에서 체크 박스를 찾고 체크 유무를 판단 할 수 있는 처리방법 제시 및 구현
* 이미지에서 테이블의 구조를 추출하고 연관관계를 도출할 수 있는 방법을 제시 및 구현

분석 프로세스 구조

* 사전작업(OCR 체크박스 부분에서 어떤 부분을 나누고 분석을 해야 하는지 점검, 그리고 진행하는 분석 환경에 대한 조사)
* 진행과정 (사전작업에서 계획한 내용을 실행하는 단계)
* 학습 데이터 특성(각 부분에 대해서 어떤 특성을 가지고 있는지)
* 전처리 작업(특성을 가진 데이터에 대해서 어떻게 처리를 해야 하는지)
* 모델 선정(처리된 데이터에 대해서 어떤 모델을 하는 것이 적합할지), 평가(선정한 모델에 대한 정확도 및 필요한 지표에 대해서 평가)
* 피드백(높은 지표 점수를 얻기 위한 방법에 대해서 고민)
* 결과(최종 출력된 모델에 대한 성과 및 적용 프로세스 설계)

1. 진행할 예제 : 예제 1의 금융거래 목적 확인서 처리( ~ 3분 30초)

* 처음 원본 PDF를 가져와서 이것을 Image로 처리할 수 있게끔 제작(포맷 변환).
* 색을 흑백으로 변환 후 각각의 색 반전을 일으켜서 사진을 처리합니다.
* 그 후 수직/수평의 값을 추출하고, 각각의 부분을 조합하고 얻어진 윤곽선들의 조합을 정렬하여 좌표 값을 이용하여 원하는 부분의 값의 부분을 추출 할 수 있도록 좌표를 구합니다.
* 최종적으로는 우리가 원하는 체크박스에 대한 부분을 크기에 따라서 가지고 올 수 있도록 합니다.

1. 사전작업 ( ~ 4분)

* 사전작업에서 각각의 금융거래 목적 확인서에서 수집할 부분을 4가지 부분으로 나눕니다.
  + 금융거래 목적의 5가지 체크박스
  + 고객확인사항의 10가지 체크박스
  + 공통부분의 예/아니요 로 이루어진 6가지 체크박스
  + 통장 양도와 관련된 4가지 체크박스

1. 진행과정 (학습 데이터 특성) ( ~ 5분)

* 이전에 진행한 사전작업을 바탕으로 모델에 학습할 데이터들의 특성을 파악합니다.
* 4가지의 특성으로 구분
  + 금융거래목적 : 5개의 체크박스 중 한 개를 검출
  + 고객확인사항 : 10개의 선택지의 체크박스 중 선택한 다수의 값 출력
  + 공통 : 체크박스와 체크 후 수정한 부분 출력
  + 고객확인 : 다양한 방식의 체크박스를 검출

1. 전처리작업( ~ 6분)

* 학습데이터 특성을 파악하면 다음 전처리작업을 진행합니다.

1. Augmentation : 데이터의 수가 적기 때문에 추출환경의 변수에 따라 데이터를 수직/수평으로 랜덤 값을 넣고 10장씩 늘려줍니다.
2. 각 데이터 특성에 따라 4가지 부분과 타겟으로 대한 정보를 별도로 저장합니다.
3. 모델 선정( ~ 7분)

* 각각 전처리한 데이터를 어떻게 처리하는가에 따라서 모델을 선정합니다.
* 전처리된 이미지를 RGB로 리스트로 변환하여 직렬화 시키는 것으로 만들었기 때문에 사용한 모델은 LogisticRegression으로 진행하였습니다.
* 또한 CNN방식으로 하는 픽처 엔지니어링으로 진행도 가능하고, 별도로 작업된 체크박스 데이터가 있다면 객체 검출방식으로도 진행할 수 있습니다.

1. 평가 및 피드백( ~ 8분)

* 이러한 방식으로 모델을 선정하면, 선정된 모델을 환경에 따라 별도로 추출한 밸리데이션 데이터나 테스트데이터로 평가를 진행합니다.
* 제가 진행한 LogisticRegression은 각각 0.95, 0.98, 0.83, 0.98의 정확도를 나타낸 것을 볼 수 있습니다.
* 이렇게 평가된 모델은 데이터를 Augmentation을 해주던가, 트레이닝 / 테스트 데이터의 비중을 다시 조절하여 설정해주거나 원천 데이터의 개수를 늘려주는 것이 대표적이지만, 환경에 따라서 1,2번만 진행해줍니다.

1. 결과 ( ~ 9분 )

* 최종 피드백까지 적용된 모델 평가 내용입니다. 물론 이것은 실제 원천데이터가 한 장을 기준으로 진행한 것이고, 많은 트레이닝 데이터로 진행한 것이 아니기 때문에 감안해주시면 감사하겠습니다.
* 이러한 모델을 적용하여 만든 아키텍처입니다. 여기서 주목해야할 부분은 실시간으로 이미지가 들어가는 부분과 모델 생성 부분이 나뉘어서 람다 아키텍처로 되어있다는 것 보시면 될 것 같습니다. 이렇게 생성된 모델은 사용자에게 빠르게 전달하여 구분된 출력결과를 확인 할 수 있습니다.

1. 공통 과제(~ 12분)

* 진행했던 프로젝트 경력에 대해서 기술 하겠습니다.
* NDS에서 진행한 마이데이터 / 라이프월릿 실증사업에 대한 내용
  + 데이터 : 별도 수집한 식품 이미지(촬영 이미지)
  + 모델 : YOLOv4 및 NCP를 사용
  + 성과 : 각 이미지에 대한 멀티 디텍팅 및 정확도 / 민감도 80%이상 달성
* 지금컴퍼니에서 진행한 인공지능 스마트 홈 프로젝트
  + 데이터 : 실제 디바이스에서 수집된 상태 정보 및 제어명령 로그
  + 모델 : 클러스터링 모델과 온프레미스 환경을 이용
  + 성과 : 설정된 데이터를 기준으로 클러스터링으로 사용자 기반 플랫폼 개발 / IBK 창공 마포 3기 선정
* 성균관대 학교 석사 졸업프로젝트(식물 질병 예측 및 사물인식)
  + 데이터 : 캐글의 Plant Pathology 2020
  + 모델 : DenseNet, EfficientNet / Augmentation : Albumentations을 코랩으로 개발
  + 성과 : 정확도를 98.06%를 기록하였고 실제로 캐클 순위 8위 기록