**연구결과서**

부서명 : 통로이미지㈜ 디지털 연구소

작성자 : 김근수

작성일: 2020.12.28

참가자: 이재봉, 천길성, 홍승호, 김근수

개발기간 : 2020년 7월 ~ 12월

|  |  |
| --- | --- |
| 연구명 | 머신러닝 이미지 텍스트 추출 |
| 결과요약 | 이미지 내의 텍스트 인식률 향상  이미지 내의 텍스트(다국어) 추출  이미지 텍스트 인식률 향상을 위한 머신러닝 |
| 결과상세 | 회사에 입고 받는 콘텐츠 중 텍스트가 없는 콘텐츠 선별을 위해  이미지내의 텍스트 추출을 위한 여러가지 기술을 검토함  광학 문자 인식(OCR)의 엔진 중 가장 대표적이며 Google에서 지원받는 Tesseract로 선정 함  Tesseract로 이미지내의 텍스트를 추출하는 것이 가능하였으나  인식률이 낮아 하단의 방법적용함  1. grayscale, thresholding, remove\_nose 등등의 이미지 선보정하여  텍스트 인식에 불필요한 영역 제거 시 인식률이 향상되는것을 확인함 2. 이미지내의 글자영역들을 그룹화하여 각 그룹별의 영역을 별도로 인지하도록 하여 특정영역의 글자에만 집중하도록 함  기본적으로 영어만 추출되어 원하는 언어의 LSTM 모델을 다운받아 적용시 다국어도 인식하게 할수 있음을 확인함  다국어 인식에만 그치지 않고 한국어의 쌍모음 인식률을 높이기 위해 JtessBoxEditor을 활용한 딥러닝(LSTM) 학습을 시도 함  하지만 쌍모음 학습한 결과 비정상동작하는것을 확인함 |
| 개선방안 | 텍스트 인식률 향상을 위한 보정작업에는 여러가지가 있으나 이미지별로 인식률 향상에 영향을 주는 보정작업이 다른것으로 확인되었다.  더 많은 이미지의 텍스트를 추출해 가며 이미지별로 최적의 보정작업을 선정이 필요함.  딥러닝(LSTM)의 현주소를 좀더 알아보고, 추후 더 자세히 알아본 후 Tesseract Command 명령어 사용 방법으로 재시도 필요함. |