



•

•

•





• **이미지 수집 출처 : 무신사 스토어 “https://store.musinsa.com/app/”**

**영수증, 포장박스 등 비정상적 이미지**

**<크롤링 된 이미지 폴더>**

****

**<데이터 수집 과정에서 발생한 문제점>**

•**‘반스-Canvas’ 카테고리의 신발이 압도적으로 많음**

• **다리 나온 사진, 영수증 사진, 신발의 일부만 나온 사진 등 불필요한 사진 존재**

**<해결 방안>**

• **데이터의 개수가 적은 카테고리의 이미지 -> Generator를 이용해서 Image Augmentation 기법 활용 -> 모델에 과적합됨 -> Generator 사용x**

• **정규화 방법 이용 -> Min-Max Scaler 방법 이용 -> 성능이 좋아짐 -> 사용o**

****

****

**- 사진 용량의 감소를 위해 500x500 -> 64x64 사이즈 조절**

**- 범주형 자료인 카테고리 -> 원 핫 인코딩을 통한 카테고리 라벨링 - Min-Max Scaler를 이용한 정규화 -> 원본 이미지의 픽셀값 0~255 -> 0~1 - 데이터의 shape, pixel, 차원수, 이미지 확인**

**- 이미지 데이터를 일일이 옮기기 힘드니 numpy 배열 파일로(.npy) 저장**

****

****

**- Conv2D, MaxPooling2D, Dropout, Flatten, Dense 레이어 사용하여 CNN 모델 설계 - Early Stopping 방법을 사용하여 6번의 진행동안 loss값이 갱신되지 않으면 훈련을 정지 - Accuracy, Loss 그래프 확인**

**- 성능이 가장 좋은 순간을 ModelCheckpoint를 통해 파일로(.h5) 저장**

**- 새로운 신발 이미지를 입력해 제대로 예측이 되는 지 확인**

****

**<데이터 수집>**

**무신사스토어 -> 스니커즈 & 신발 카테고리 -> 상품평순 정렬 -> 상위 20개 상품 -> 일반 후기 최대 10 페이지 크롤링**

**<수집 과정에서 발생한 문제점>**

**- 리뷰가 없는 신발 존재**

**- 리뷰 작성자가 동일한 경우 존재**

**<해결 방안>**

**- 리뷰가 없는 신발의 공식 브랜드 홈페이지에 가서 후기를 크롤링**

**- 텍스트 전처리 과정에서 중복 데이터를 제거하도록 코딩**

**<분석 과정>**

**1. 중복 데이터, 결측값, 특수기호, 이모티콘, 숫자 제거 (한글, 띄어쓰기만 남기기) 2. 띄어쓰기, 맞춤법 교정**

**3. 토큰화**

• **Mecab, Okt**

• **다른 라이브러리를 사용하지 않은 이유 : 속도가 너무 느리고 성능도 Mecab, Okt와 비슷하기 때문**

• **일반적인 토큰화 방법을 사용하지 않고, 각 라이브러리의 품사 홈페이지에 들어가서 명사, 형용사, 부사만 토큰화 하도록 품사의 형태를 지정 4. 불용어 제거**

**5. 토큰의 최대 길이 정하기**

**6. 단어 사전 만들기**

**- 단어의 빈도수 높은 순으로 나열**

**- 제일 많이 사용된 단어 10개를 해당 카테고리(‘브랜드-종류’)의 추천 태그로 사용**

**7. 최종 데이터프레임에 카테고리와 추천 태그를 미리 넣어놓기**

****

**- 새로운 테스트 이미지를 함수에 집어넣게 되면, CNN 모델을 통해 해당 신발의 카테고리(‘브랜드-종류’)를 예측하고, 그 예측결과 에 맞는 추천태그를 데이터프레임에서 찾아서 결과적으로 신발에 해당되는 카테고리와 추천태그를 출력**

**- 이미지 CNN 모델링**

• **메인 사진 외에 세부 사진비율이 더 많아서 예측율과 Loss가 떨어지는 결과가 발생**

• **추가적으로 전이 학습 방법 이용 -> 다른 사람들이 미리 학습해 놓은 모델 이용 -> 모델의 정확도가 높아지기를 기대 -> 오 히려 이전보다 성능이 안 좋게 나옴 -> 전이 학습 방법을 사용x**

**- 텍스트 전처리**

• **두 가지 라이브러리를 사용했는데 Okt가 우리가 원하는 결과에 더 가까운 글자의 형태를 나타내므로 Okt를 사용함** • **일반적인 불용어 뿐만 아니라, 신발을 분석하는 것이니까 신발의 브랜드 이름, 신발의 종류, 기타 불용어들을 더 추가함**