내이버 쇼핑
상품 이미지 유사도 추출 과제

작성자 : 이성철
소속팀 / 상위부서 : 쇼핑데이터개발/ Forest CIC, 비즈OCR / Glace CIC

일반

NAVER

목차

D2 CAMPUS FEST 쇼핑데이터 플랫폼 과제

과제명 : 이미지 유사도 측정하기

데이터셋 : 쇼핑 이미지 제공(1만장 가량 제공예정)

평가 : 쇼핑데이터 플랫폼에서 정답셋 제공. 정답과의 오차율이 적을 수록 우수

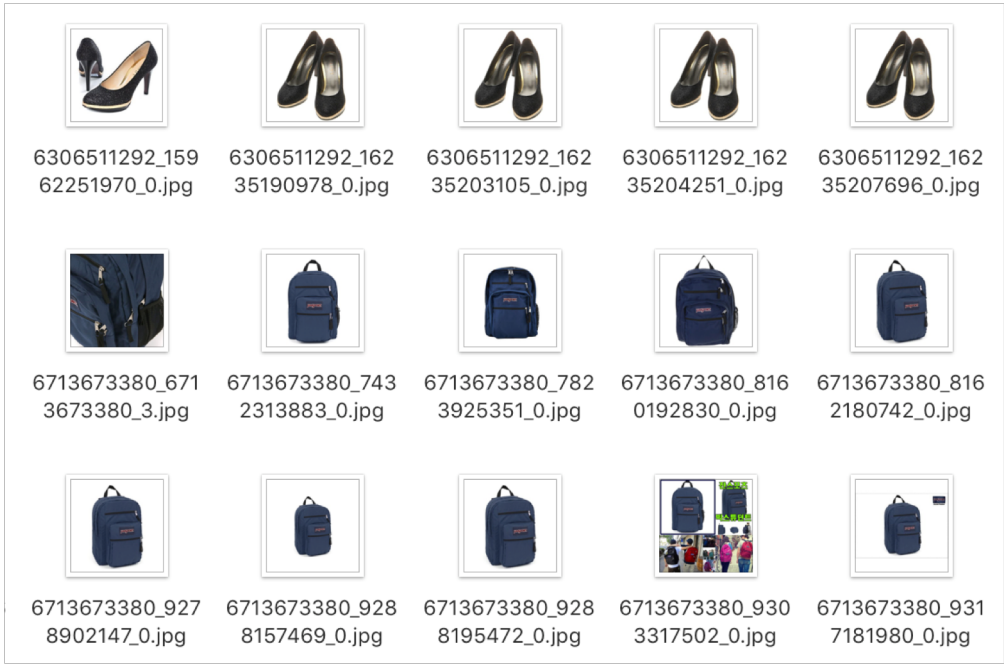
과제 진행

이미지가 가진 고유한 특징점들을 AI/ImageProcessing등을 활용하여 추출

몽고DB/메모리/기타/DB/text file등을 활용해서 유사도 ThresHold 95%이상 유사한 이미지들을 group화하여 제출(clustering)

1.1 데이터셋 설명

	Training		Test (미제공)	
카테고리	모델수	상품수	모델수	상품수
가방	79	5320	18	1270
여성구두	62	5202	16	1417
총	141	10522	32	2687



2. Baseline 코드 설명

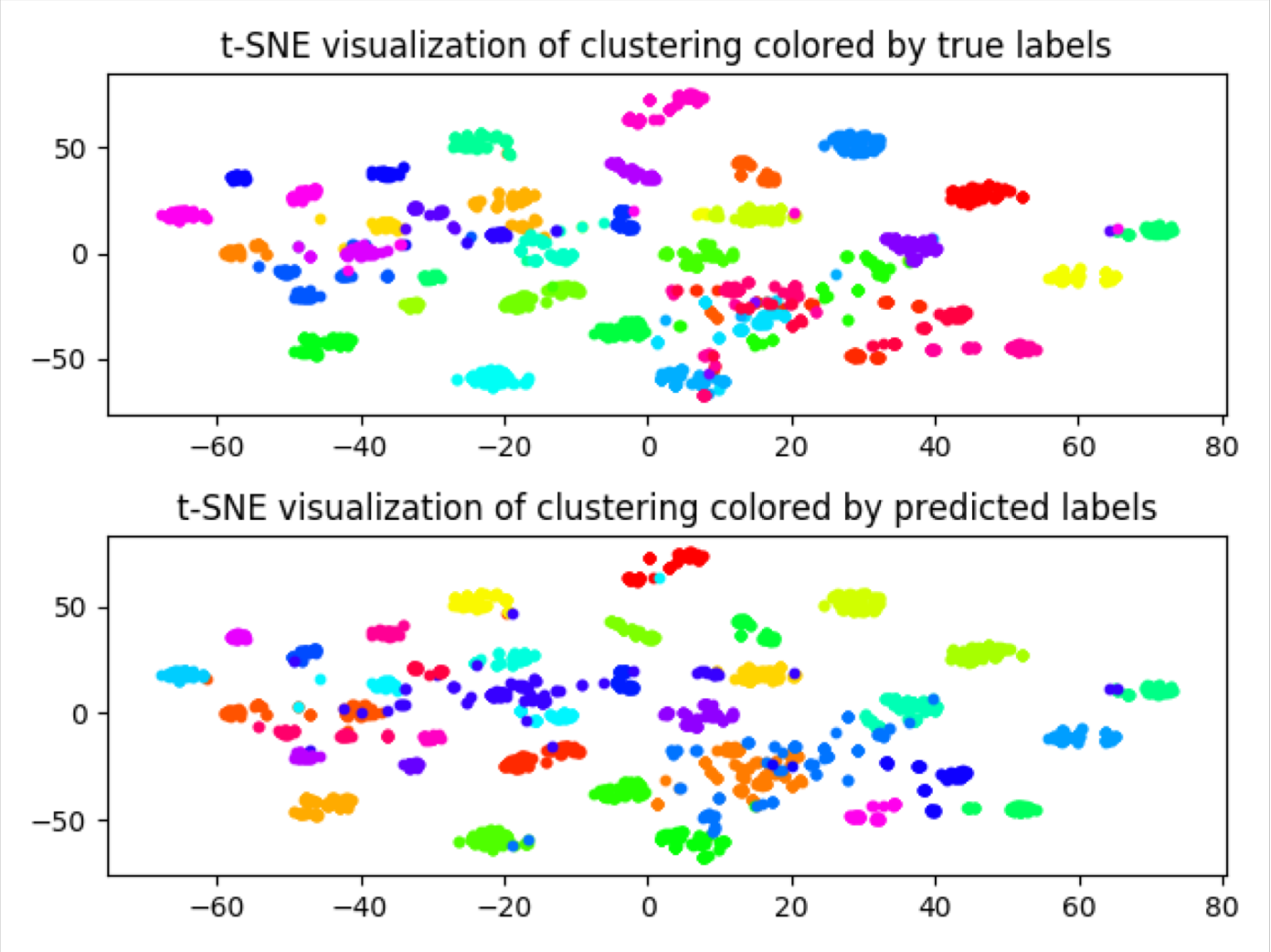
1. make_labels_true.py
   - 클러스터링을 평가하기 위해 파일명-> 클러스터 라벨 변환
   - labels_true.npy 에 저장
2. extract_features.py
   - MobileNet V2를 사용하여 특징 추출 (pre-trained model from TensorFlow Hub)
   - 추출된 특징을 features.npy 에 저장
3. make_labels_pred.py
   - 특징을 로딩하고 클러스터 개수를 예측 함
   - K-means 알고리즘으로 클로스터링 함 결과 -> labels_pred.npy
   - labels_true 와 labels_pred 를 비교하여 클러스터링 점수 (Adjusted Rand index) 계산

```
Estimated num_clusters: 35
0.758350016511794
```

4. visualization.py
   - T-SNE로 클러스터링 결과를 시각화 할 수 있음
   - .tsv 파일을 http://projector.tensorflow.org/ 에 업로드하면 인터랙티브한 시각화가 가능

Adjusted Rand index: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

3. 시각화


t-SNE visualization of clustering colored by true labels
50
0
−50
−60
−40
−20
0
20
40
60
80
t-SNE visualization of clustering colored by predicted labels
50
0
−50
−60
−40
−20
0
20
40
60
80

4. 힌트

Feature extraction, distance metric, clustering 등 다양한 면에서 개선 할 수 있을 것임

1. Feature extraction
   - Training set으로 pre-train 모델을 fine-tuning 함
2. Distance metric learning
   - http://sanghyukchun.github.io/37/
3. Clustering
   - 키타 클러스터링 알고리즘을 조사하고 적용함

참조

https://github.com/EdjoLabs/image-match

https://www.tensorflow.org/hub/

http://projector.tensorflow.org/

https://scikit-learn.org/stable/modules/clustering.html

https://towardsdatascience.com/unsupervised-learning-with-python-173c51dc7f03

End of Document

Thank You.