과학기술정보통신부 NIA 지능정보원 한국지능정보사회진흥원

DATA CREATOR CAMP

2024 데이터 크리에이터 캠프

대학부 팀명 : Data Nexus

1-1 성별&스타일 통계표 작성

해석: 이미지 파일명을 이용해 Training 및 Validation 데이터에서 "성별 & 스타일" 통계를 도출하여 분석.

데이터 로드 및 성별/스타일 추출 : 파일명을 기반으로 이미지ID, 성별, 스타일정보를 추출
예: W_65122_10_sportivecasual_W.jpg➔ 성별: W(여성), 스타일: sportivecasual

분류 과정: 추출된 성별 & 스타일별로 데이터를 Training, Validation 세트로 나누어 분류
count_images함수: Training 및 Validation 데이터에서 각 성별/스타일별 이미지 수를 카운트

각 데이터 세트에서 성별 & 스타일별 이미지 수를 집계하여 통계표 작성
->dict_to_dataframe함수로 성별 & 스타일별 이미지 수를 표 형식으로 변환
->reshape_dataframe함수로 데이터를 가독성 높은 형식으로 재구성

최종 통계 출력
재구성된 표는Training과Validation데이터에서 성별과 스타일별 이미지 분포를 요약하여 시각화

1-1 성별&스타일 통계표 작성

Training 데이터

성별	스타일	이미지 수
여성	feminine	154
여성	normcore	153
여성	sportivecasual	157
남성	normcore	364
남성	hiphop	274
...	...	...

Validation 데이터

성별	스타일	이미지 수
여성	feminine	44
여성	minimal	35
여성	sportivecasual	48
남성	normcore	51
남성	hippie	82
...	...	...

1-2 ResNet-18을 활용한 성별 & 스타일 분류

데이터 준비 → 데이터 전처리 → 모델 구성 → 학습 및 검증 → 모델 저장

<훈련 및 검증 데이터 전처리 과정 정리>

과정	훈련 데이터 전처리	검증 데이터 전처리
크기조정	224x224 픽셀	224x224 픽셀
랜덤수평 뒤집기	100% 확률로 적용	미적용
랜덤회전	-10~10도 사이에서 무작위 회전	미적용
색상조정	밝기, 대비, 채도, 색조 변환	미적용
정규화	Mean=[0.4933, 0.4610, 0.4464], Std=[0.2573, 0.2508, 0.2519]	동일하게 적용

1-2 ResNet-18을 활용한 성별 & 스타일 분류

<ResNet-18 모델 구성>

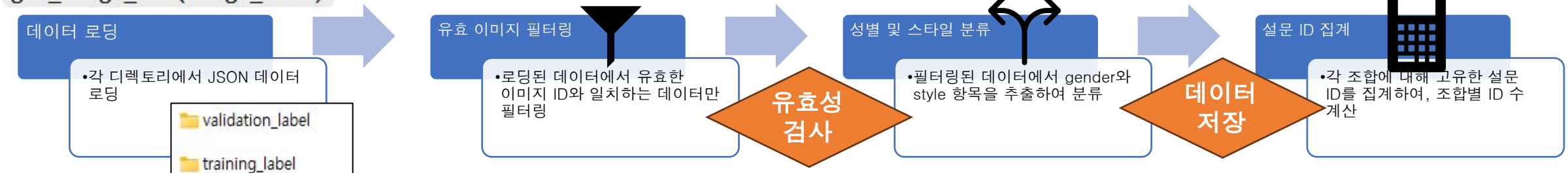
<학습 및 검증 설정 요약>

항목	방법
손실함수	CrossEntropyLoss(클래스 가중치 적용)
최적화 기법	Adam Optimizer, learning rate = 0.01, weight_decay=1e-3
학습률 스케줄러	ReduceLROnPlateau(optimizer, mode='min', factor=0.5, patience=5)
평가지표	Top-1 accuracy (Validation 데이터 기준)
총 Epoch 수	300회 학습
모델 저장 경로	resnet18_fashion_classification.pth

2-1 "성별&스타일"에 따른 설문ID 기반 통계

`{W/T}_{이미지ID}_{시대}_{스타일}_{성별}_{설문ID}.json`

`get_image_ids(image_dirs)`



`collect_label_info_with_respondent(label_dirs, image_ids, dataset_name, image_dir)`

Training 데이터

스타일	남성	여성
athleisure	43	381
bodyconscious	48	452
bold	133	1011
cityglam	96	233
classic	143	361
…	…	…

Validation 데이터

스타일	남성	여성
athleisure	0	80
bodyconscious	3	111
bold	6	215
cityglam	6	59
classic	19	98
…	…	…

Feedback : 문제에 대한 정확한 이해, 매칭 해결, 유효한 데이터 도출

결론 : 남성보다 여성의 특정 스타일에 설문이 집중된 경향이 있어, 모델 학습 시 성별 편향이 발생할 가능성 있음

2-2 유효 라벨링 데이터 중 상위 100명 응답자 선호도 분석

`combined_label_info_with_resp`

	respondent_id	train_스타일선호	train_스타일비선호	valid_스타일선호	valid_스타일비선호
1	64747	enderless_W_158671.jpg	0_grunge_W_158643.jpg	powersuit_W_158675.jpg	70_hippie_W_158674.jpg
2	64346	etrosexual_M_147916.jpg	normcore_M_147919.jpg	0_hiphop_M_147890.jpg	normcore_M_147894.jpg
3	65139	0_hiphop_M_174231.jpg	14_50_ivy_M_174210.jpg	tivecasual_M_174233.jpg	etrosexual_M_174240.jpg
4	64561	0_oriental_W_152840.jpg	_70_punk_W_152837.jpg	tivecasual_W_152857.jpg	0_minimal_W_152877.jpg
5	63405	_60_mods_M_122233.jpg	normcore_M_122223.jpg	94_50_ivy_M_122232.jpg	0_hiphop_M_122211.jpg
6	63369	tivecasual_M_120837.jpg	normcore_M_120830.jpg	12_50_ivy_M_120814.jpg	1_80_bold_M_120809.jpg
7	59704	70_hippie_M_087695.jpg	1_80_bold_M_087696.jpg	70_hippie_M_087687.jpg	33_50_ivy_M_087668.jpg
8	64633	4_80_bold_M_154754.jpg	etrosexual_M_154756.jpg	4_80_bold_M_154754.jpg	70_hippie_M_154712.jpg
...	...	... 2	... 1	... 2	... 1

Feedback : 설문id와 응답자id 혼동

데이터 필드:

- 응답자 ID: JSON의 USER > R_id
- 스타일 선호 여부: JSON의 item > survey > Q5 (1:비선호, 2:선호)

결과물: 상위 100명 응답자의 스타일 선호 정보를 별도의 JSON 포맷으로 저장

3-1 협업 필터링 user-based filtering, item-based filtering 비교

- ***User-based Filtering (사용자 기반 필터링)*** : 비슷한 선호도를 가진 응답자들이 특정 스타일을 선호/ 비선호 하는지를 바탕으로 새로운 응답자가 이들을 선호/ 비선호 스타일을 추천해주는 방식

적용 방법: 각 응답자의 스타일 선호/ 비선호 데이터를 바탕으로 응답자 간 유사도 계산 (유사도는 코사인 또는 피어슨 상관계수 사용) -> 벡터로 변환하여, 이를 통해 유사한 응답자를 찾는 방법

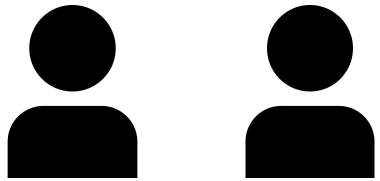
vs

⚫ ***Item-based Filtering (아이템 기반 필터링)***: 응답자가 선호하는 스타일과 유사한 스타일을 추천하는 방식으로, 선호했던 스타일 정보를 바탕으로 유사함을 예측한다.

적용 방법: 응답자의 스타일 선호 데이터를 바탕으로 응답자의 각 선호 스타일 간의 유사도를 측정 (유사도는 코사인 또는 피어슨 상관계수 사용) -> 벡터로 변환하여, 이를 통해 유사한 스타일을 찾는 방법

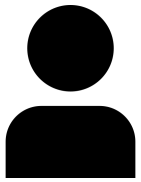
3-1 협업 필터링 user-based filtering, item-based filtering 비교

[예시-가정]



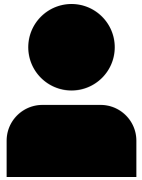
ID: 64747 ID:64633

선호하는 스타일 유사
서로 관계의 유사도 $(가정)
'hippie' 스타일 선호



ID: 69704
위 응답자와의 유사도 $
'hippie' 스타일 선호할 것!

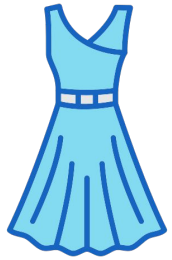
vs



ID:64633

'hiphop' 스타일 선호

서로 관계의 유사도 @(가정)
높음!



'minimal' 스타일 선호

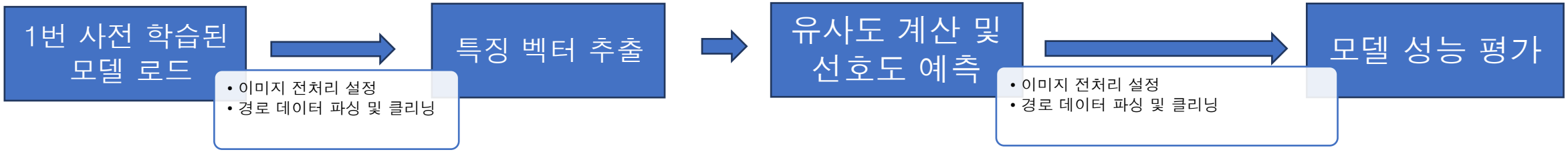
ID:64633 'minimal' 스타일 선호할 것!

3-1 협업 필터링 user-based filtering, item-based filtering 비교

[장단점 비교]

	user-based filtering	item-based filtering
장점	– 직관적 – 개인화된 추천 가능 – 선호하는 스타일 비슷한 응답자가 있을 경우 매우 효과적!	– 안정적, 다양한 추천 제공 – 데이터 부족 문제 해결에 유리 – 응답자 스타일 고정적 => 계산 비용 낮음
단점	– 데이터 부족 ⇒ 유사도 측정 어려움 – 많은 응답자에 대한 유사도를 계산=> 계산 비용이 높음.	– 응답자 개별 취향을 충분히 반영 못할 가능성 O – 특정 스타일에 대한 선호가 강한 응답자의 특이성을 반영 못할 가능성 O => 맞춤 추천이 제한적

3-2 item-based filtering 구현 후 성능 측정



< 유사도 계산 및 선호도 예측 정리 >

과정	훈련 데이터 전처리
특징 벡터 추출	응답자별 모든 이미지의 특징 벡터 추출
유사도 임계값 설정	유사도 기준 0.7로 설정
응답자별 데이터 수집	응답자 ID, 선호/비선호 학습 및 검증 이미지
유사도 계산	검증 이미지 <-> 학습 이미지 간 코사인 유사도 계싼
선호도 예측	학습 선호 이미지와의 최고 유사도 > 비선호 이미지와의 유사도/ 임계값 => '선호' 예측

3-2 item-based filtering 구현 후 성능 측정

1번 사전 학습된 모델 로드 → 특징 벡터 추출 → 유사도 계산 및 선호도 예측 → 모델 성능 평가

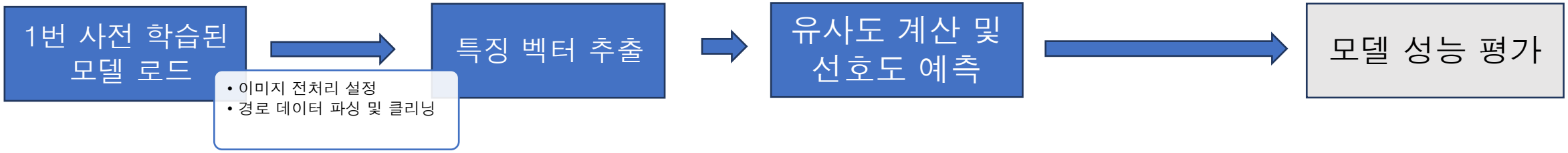
- 이미지 전처리 설정
- 경로 데이터 파싱 및 클리닝

ident_id	valid_image	predicted_preference	actual_preference
64747	dataset/validation_image/W_38588_19_genderless...	True	True
64747	dataset/validation_image/W_44330_10_sportiveca...	True	True
64747	dataset/validation_image/W_37491_70_military_W...	True	True
64747	dataset/validation_image/W_22510_80_powersuit_...	True	True
64747	dataset/validation_image/W_30988_90_kitsch_W_1...	False	True
...	...	...	...
64295	dataset/validation_image/W_32314_19_normcore_M...	True	False
64295	dataset/validation_image/W_25761_90_hiphop_M_1...	False	False
64295	dataset/validation_image/W_31478_19_normcore_M...	False	False
64295	dataset/validation_image/W_16374_10_sportiveca...	False	False
64295	dataset/validation_image/W_23900_50_ivy_M_1466...	False	False

< 유사도 계산 및 선호도 예측 정리 >

과정	훈련 데이터 전처리
특징 벡터 추출	응답자별 모든 이미지의 특징 벡터 추출
유사도 임계값 설정	유사도 기준 0.7로 설정
응답자별 데이터 수집	응답자 ID, 선호/비선호 학습 및 검증 이미지
유사도 계산	검증 이미지 <-> 학습 이미지 간 코사인 유사도 계산
선호도 예측	학습 선호 이미지와의 최고 유사도 > 비선호 이미지와의 유사도/ 임계값 => '선호' 예측

3-2 item-based filtering 구현 후 성능 측정



	respondent_id	precision (정밀도)	recall (재현율)	f1_score (F1 점수)
0	368	1.000000	0.833333	0.909091
1	837	1.000000	0.600000	0.750000
2	7658	1.000000	0.333333	0.500000
3	7905	1.000000	0.500000	0.666667
4	9096	0.666667	0.400000	0.500000
...	...	...	...	...

결론: 모델이 예측할 때 비교적 신뢰할 수 있는 결과를 제공

‘Data Nexus’ 팀원들 소감

구동한

미션을 수행하면서 가장 먼저 문제를 명확하게 이해하는 게 중요한 것 같다는 생각이 많이 들었습니다.
어떤 방식으로 어떤 기준을 가지고 문제를 해결해야 할지 모호한 경우가 있어. 수행하는데 생각보다 시간이 많이 소요되었고 오류를 해결하고 테스트 하는 과정에서도 많은 시간이 들었습니다. 막바지에 좀 더 여유가 있었으면 했지만 짧은 시간 동안 데이터 처리 하는 방법에 대해 많이 배운 것 같아 뜻 깊은 시간이었습니다.

노을

캠프를 참여한 당시엔 실제로 미션을 받아 데이터 분석을 해본 적은 처음이었습니다. 직접 문제를 파악하고 데이터를 분석하기 위해 생각해보는 시간을 가지며 차차 성장해 나가고 있음을 느낄 수 있었습니다! 데이터 분석에 있어서 문제를 잘 파악하고, 어떤 데이터를 전처리하고 활용하며 분석을 해야 하는지가 가장 중요한 것 같다고 생각했습니다.

정유진

데이터 수집과 해석 능력이 중요해지는 요즘, 데이터를 다루며 그 과정의 복잡함을 깨달았습니다. 여러 도전에 직면하면서 팀과의 협업 가치를 재발견하고 최적의 해결책을 찾는 과정은 제 성장에 큰 도움이 되었습니다. 또한, 협업 필터링 기법을 통해 사용자 선호를 분석하고 시각적으로 표현하는 경험은 타 연구에서도 유용하여, 이 계기로 앞으로도 계속 배워 나가겠습니다.

홍다빈

데이터 크리에이터 캠프에 참여하면서 처음 접하는 내용들이 많아 예상보다 어려움을 느꼈습니다. 이론만으로는 부족하다는 것을 깨달았고, 실제로 다양한 실습을 통해 여러 시행착오를 겪으며 제 부족함을 더 많이 느꼈습니다. 이러한 경험을 통해 앞으로 더 열심히 공부 해야겠다는 결심을 하게 되었습니다.

DATA CREATOR CAMP

2024 데이터 크리에이터 캠프

감 사 합 니 다

과학기술정보통신부 NIA 지능정보원