슬라이드1

안녕하세요 25년 경영학과 컴퓨터공학 복수전공중인 김지수입니다. 저는 소비자 구매 행동 예측을 위한 Multimodal 데이터를 활용한 모델에 대하여 발표하려고 합니다.

슬라이드2

사실 소비자의 구매, 의사결정과 관련된 연구는 경영학에서 오래 전부터 소비자 행동론이라는 분야에서 연구가 이루어졌습니다. 대표적인 키워드로는 의사결정에 있어 휴리스틱, 프레이밍효과, 확인편향, 불확실성하의 의사결정 등이 있습니다. 관련 수업으로는 23년 1학기 하영원 교수님의 수업을 수강하였습니다.

슬라이드3

한편 아시다시피 세계적으로 커머스는 온라인으로 비중이 늘어났으며 국내에서도 거래액수를 기준으로도 많은 항목에서 절반 가까이 되고 있습니다.
여기서 과거부터 소비자의 의사결정에는 국적 등의 문화적 특성을 나타내는 변수의 차이가 의사결정 스타일에 미치는 영향에 대하여 지속적으로 연구가 되어 왔습니다. 구체적으로 문화적 특성에 따라서 소비자의 리스크 회피 성향, No1 브랜드 선호성향, 할인에 대한 반응, 무료 배송 여부에 대한 반응 등이 다르다는 것입니다. 저는 이 지점에서 호기심을 가지고 주제를 고민해보았습니다.

슬라이드4

그렇다면 이러한 연구를 위해서 멀티모달 데이터를 활용한 온라인 환경에서의 소비자 구매와 관련하여서는 인공지능을 기반으로 다음과 같은 연구가 진행되었습니다. 온라인 커머스의 상품 이미지, 상품 설명 생성, 상품 이름 번역, 상품 추천 모델, 구매 예측, 만족도 조사, 상품 분류 등과 같이 폭 넓은 관련 분야에서 다방면으로 연구가 이루어지고 있습니다.
추천 시스템은 주로 상품 이미지와 리뷰 텍스트를 기반으로, 예측의 경우에는 로그데이터를 기반으로 하고 있으며 네이버 쇼핑에서는 상품의 클러스터링, 속성 추출, 카테고리 분류, 유해 상품 인식 등에 이를 사용하고 있었습니다.

슬라이드5

추가적으로 데이터를 활용해서 군중 행동을 예측하거나, 성격 특성을 예측하는 연구는 진행되었으나 소비자 의사결정과 관련하여 상품 텍스트 설명 생성, 상품 썸네일 선택 등 구매에 영향을 줄 수 있는 항목에 대하여 해당 선택을 하는 소비자의 문화적 특성을 직접적으로 반영한 연구를 찾지는 못했습니다. 한편, Amazon 데이터 셋의 경우 소비자의 국적은 간접적으로 데이터를 사용하고 있고 6개국의 데이터만 포함되어 있는 것처럼 해당 연구는 산학협력이 이루어졌을 때 훨씬 가치가 높을 것입니다.

슬라이드6

이러한 배경지식을 바탕으로 멀티모달 데이터를 활용한 소비자 구매 행동 예측 문제를 정의했습니다. 데이터로는 추천, 군집화, 설명 텍스트 생성, 예측 등 개발하려는 모델의 용도에 따라서 다음의 데이터들 중에서 선택할 수 있을 것입니다. 모델로는 기본적으로 여러 형태의 데이터를 처리할 수 있는 트렌스포머 모델을 사용하면서, 의미 있는 분석을 위하여 attention weight 분석이나 기여도 분석을 실행할 수 있을 것입니다.

특히 attention weight분석으로 모델이 각 입력 요소에 부여한 중요도를 살펴봄으로써 만약 한국 소비자용 모델이 한정 수량과 같은 단어에 더 높은 가중치를 부여하고 미국 소비자용 모델이 무료배송과 같은 문구에 더 높은 가중치를 두었다면 실제로 커머스 플랫폼에서 사용 모델을 선택함에 있어 소비자의 문화적 특성을 반영하는 것이 구매전환 확률에 영향을 줄 수 있을 것이라고 생각했습니다. 단순히 정확도를 높이는 것에 더해서 어떤 요인이 어떤 modality에서 영향을 미치는지 구조적인 설명이 가능하다면 이는 향후 소비자 행동에서 확장해서 다국적 기업의 추천 시스템, 글로벌 광고, 지역화된 UX 디자인 설계 등에서도 활용 가치가 있다고 생각합니다.

슬라이드 7

연구 질문은 다음과 같습니다. 해당 분야에서 다양한 연구가 이루어질 수 있어서 여러 가지를 작성해보았습니다만 각 연구 질문이 별도의 연구가 필요할 수 있습니다.

우선, 문화적 특성을 반영한 Multimodal 데이터의 사용이 기존 모델보다 성능적으로 좋은지, 좋다면 어떤 데이터 요인에 의해 더 설명되는지 분석할 수 있다면 매우 좋겠습니다. 추가적으로 로컬 모델과 글로벌 추천 모델이 소비자 구매 의사결정에 어느 정도의 영향을 줄 수 있을지 사후 추적한다면, 데이터 부족 문제의 상황에서 모델을 확장할 때 기업의 의사결정에 참고가 될 수 있을 것이라고 생각했습니다.

슬라이드 8

다음으로 멀티모달을 활용한 모델을 사용하여 Kaggle에서 우수한 성적을 거둔 케이스를 소개하고자 합니다. 찾아본 결과, 기존 CV라고 하는 이미지 분야에서 대체적으로 Multimodal을 사용한다고 해서 모델 성능이 딥러닝 이미지 처리 모델들보다 좋은 것은 아니였습니다. 다만, 여기서 소개드리는 케이스와 같이 특정 사례에서는 Multimodal 기반의 모델의 성능이 우수하게 나타났습니다. 개인적인 생각으로는 이미지의 경우 이미지 자체만으로 높은 성능이 얻어질 수 있는데 앞으로 경영학과로서 기업에서 다루는 복잡한 문제들의 경우에는 multimodal을 잘 활용하면 좋은 성능을 얻을 수 있지 않을까 생각해보았습니다.

대회는 Shopee라는 동남아시아를 주된 타겟으로 하는 쇼핑 플랫폼으로 이미지와 텍스트 데이터를 기반으로 제품이 동일한 제품인지를 분류하는 모델을 만드는 것이 목적인 대회입니다. 이는 온라인 쇼핑시에 같은 제품이라도 가격이 다른 경우가 많기 때문에 상품을 제시할 때 제품을 카테고리화하고 스팸 상품을 가리는 목적이나 정확한 유사상품 리스팅을 통해 최적의 Deal을 제안함으로써 소비자 경험의 향상을 위해 사용됩니다.

슬라이드 9

From Embeddings to Matches

1. 모델 예측 성능 정리

- **Baseline(기본)**: 0.70 (이미지만 사용), 0.64 (텍스트만 사용)
- 이미지 임베딩 + 텍스트 임베딩 **concate 후 정규화**: 0.724
- **min2 전략 적용**: 0.743
- **각각을 정규화 후 concat** (순서 변경): 0.753
- **전체 데이터로 학습**: 0.757
- **이미지/텍스트 매칭 결과 합친후 임계값 튜닝**: 0.776
- **INB 기법 + 다양한 텍스트 모델 추가**: 0.784
- **INB 1단계에서 이미지/텍스트/조합 임베딩 모두 활용 + 임계값 공동 튜닝**: 0.793

슬라이드 10

- 이미지 인코더와 텍스트 인코더를 각각 사용해서 이미지와 텍스트 데이터를 임베딩해 이를 활용하여 매칭을 수행함

이미지 인코더: timm 라이브러리의 eca_nfnetl1 모델 2개 사용
이미지 feature는 GAP(Global Average Pooling) 후
BatchNorm1D, Normalization 단계를 거쳐서 최종 1792 크기로 생성

텍스트 인코더는 Hugging Face 플랫폼에 있는 아래의 모델 사용
xlm-roberta-large, xlm-Roberta-base
cahya/bert-base-Indonesian-1.5G (인도네시아 언어)
Indobenchmark/indobert-large-p1
Bert-base-multilingual-uncased
텍스트 featur는 pooling, batchnorm1d, Normalization을 통해 1024 크기로 생성

Cosine similarity 계산 후 결과에 Arc Margine 적용 후
Softmax Loss를 사용하여 학습을 진행함
최종적으로 얻은 image/text embeddings가 kNN search에 활용되어
이미지-텍스트 간 유사도를 기반으로 매칭이 수행됨.

슬라이드11

3. ArcFace Tuning (생략)

이전 과정에서 Arc Margin 적용할 때 임베딩에 충분히 마진을 주는 것이 중요하다는 내용 (논문을 기반으로)

4. 이미지 & 텍스트 매치 혼합 - Concatenation & Union

이미지와 텍스트 두 입력을 어떻게 잘 결합하는지가 성능에 매우 중요
Image matches + Text matches + Comb maches -> Union 이 가장 좋다는 결론을 얻음

임베딩을 정규화한 이후 Concatenation하고 Comb Similarity를 계산 (이미지 유사도와 텍스트 유사도를 평균낸 것과 동일함)
이미지 기반으로 강하게 추천된 항목, 텍스트 기반으로 강하게 추천된 항목, 둘 다 중간 정도로 추천된 항목을 모두 받아들이는 전략
이미지와 텍스트 인코더를 Joint training도 시도해보았으나 성능은 개별 학습 후 합친 경우보다는 낮았음

슬라이드12

5. Iterative Neighborhood Blending (INB)

핵심 아이디어: 임베딩 간의 유사도를 기반으로 이웃을 정의하고, 해당 이웃 정보를 활용해서 임베딩을 반복적으로 개선하는 것 (더 정밀한 군집화와 매칭 향상)
INB 구성요소)

1. K-Nearest Neighbor Search (k=51)

- faiss(https://github.com/facebookresearch/faiss) 사용
- cosine similarity로 유사한 top-50 이웃 검색 (문제 제출 최대 50개)

2. Thresholding (기준치 정하기)

- cosine similarity → cosine distance(= 1 - similarity)로 변환
- 일정 threshold보다 가까운 이웃만 유지
- min2: 최소 2개의 매칭을 확보하도록 두 번째 이웃까지는 threshold를 완화

3. Neighborhood Blending (NB)

- 각 아이템의 이웃 임베딩을 **유사도 기반 가중치 합**하여 자신에게 더함
- 유사도가 edge weight인 그래프처럼 생각

→ 결과적으로 **더 조밀한 임베딩 공간 형성**

4. Iterative NB

- 위 과정을 여러 번 반복 (stage1 → stage2 → stage3)
- 각 단계에서 refined embedding 생성 후 다시 kNN + NB
- 반복은 성능이 개선될 때까지만 진행함

슬라이드13
전체적인 구조로 각각의 image/test, 그리고 둘을 혼합한 Embedding에서 maches와 유사도를 추출하고 합쳐서 Match Set을 구성합니다. 그리고 이전의 NB 이웃 임베딩을 반복하는 과정을 거쳐서 최종의 매치 결과를 생성한다고 알 수 있습니다.

슬라이드14
7. Discussion

- 이렇게 Image, Text, Comb embedding 과 threshold를 jointly tuning한 모델의 성능이 0.793으로 가장 높았음.
- product matching could support more accurate product categorization(분류) and uncover marketplace spam(이상한 제품) <- 이전에 Naver에서 Multimodal 활용한 것과 동일 Task
- posting_id,matches

test_123,test_123 <- 123번 id의 상품은 스스로만 동일

Test_456,test_456 test_789 <- 456번 id의 상품은 789랑 동일 (최대 50개까지 매칭)

- 0.780의 Private(비공개 테스트셋) F1 Score 달성 (평균적으로 78% 정확도 수준으로 매칭을 정확히 찾아냄)
- 아직 이런 Multimodal 모델에 대해서는 계속 연구가 이루어지고 있어서 추후 사용되는 분야가 넓혀지고 모델이 정교화된다면 더 좋은 성능을 기대해볼 수 있을 것 같습니다. 또 Multimodal이라는 모델이 데이터의 형태를 여러개를 사용하는데, 이제 데이터는 의도하던 의도하지 않던 매우 방대하게 자동적으로 기록되고 있기 때문에 우리가 해결하고자 하는 문제에 어떤 데이터가 효과적일지 고민하는 것도 중요하다고 생각했습니다.