Amazon Review Dataset을 활용한 Recommender System

CUAI 5기 RecSys11팀
권예진(응용통계학부), 김중훈(응용통계학부), 박경빈(소프트웨어학부), 임도연(소프트웨어학부)

[요약] Amazon Review Dataset을 이용하여 추천시스템 모델을 구현한다. 그 과정에서 자연어처리 기술을 이용하여 Meta Data의 상품 description 항목을 임베딩하여 성능 개선을 목표로 한다. 또한 그 과정에서 최신 모델인 BiVAE를 이용하여 성능 개선 효과를 얻는다.

1. 서 론

추천시스템이란 User-Item Interaction Data(사용자-아이템간의 상호작용 데이터)와 Meta Data(아이템 데이터)를 이용하여 특정 사용자에게 아이템(영화, 음악, 링크 등)을 추천해주는 기술이다. 유튜브의 일명 '알고리즘'이라 통하는 기술과 동일하며, 유튜브의 경우 여러 사람들의 시청 기록, 추천 영상 클릭 여부, 좋아요 한 항목 등을 반영하여 새로운 영상 아이템을 추천해준다. 다양한 사업에 적용될 수 있고, 또 적용되고 있기 때문에 해당 기술의 발전을 통해 사용자에게 더 좋은 환경을 제공해 줄 수 있으리라 생각하여 본 연구를 진행하게 되었다.

이런 추천 시스템 분야에서 최근 RLP (Recommendation as Language Processing)[1] 등 자연어처리 기술을 이용하는 동향이 있었다. 따라서 본 논문에서도 자연어처리 기술을 이용해 본 후, 그 결과를 자연어처리 기술을 적용하지 않은 모델과 비교해 어느 정도의 성능 향상이 있을지 알아보기로 하였다.

2. 본 론

1) 데이터셋

데이터(정의) 선택한 데이터는 Amazon 회사의 2018년 Review Data로 본 논문에서는 Grocery and Gourmet Food 항목의 리뷰 1,143,860개와 제공된 Meta Data를 이용하였다. Meta Data의 경우 제품 상세 설명에 해당하는 Description 항목과 가격인 Price 값, 그리고 해당 제품의 category를 선정, 이용하였다.

2) 전처리

전처리의 경우 Colab pro 버전의 고용량 RAM을 이용하였으나, 최적화된(optimizer) 모델을 찾기 힘든 추천시스템 기술의 특성상 한계가 생겨 불가피하게 이와 관련된 전처리를 진행하였다. 따라서 아이템을 기준으로 구매한 사람이 5명 이상, 구매한 이력이 30번 이상인 사람들의 Interaction Data만 남겨두어 총 n개의 데이터를 이용하였다. 또한 Meta Data의 category의 경우 Multi Binarizer를 이용하였는데, 이는 다중 클래스에 대한 원-핫 인코딩(one-hot encoding) 방식으로 설명될 수 있다.

3) NLP Task

메타 데이터의 description을 모델에 이용하기 위하여 sBERT 모델[2]을 사용하였다. sBERT의 임베딩 과정 중 하나인 평균 풀링을 이용하였는데 이는 BERT의 각 단어에 대한 출력 벡터들에 대해서 평균을 내고 이를 문장 벡터로 생각하는 것이다. 해당 과정에서는 문장 임베딩만을 진행하였으므로, 기존의 BERT 방식보다 문장 임베딩 성능이 더 우수하다는 장점이 있고, 각 제품 당 description의 길이가 단어 100개를 넘는 경우가 많아 훨씬 적합하므로 본 모델을 채택하였다. 최종적으로는 (6186, 768) 차원의 임베딩 값을 얻을 수 있었고 그 예시는 아래 그림과 같다.

	asin	embedding
0	4639725043	[-0.36482325196266174, 0.8115509748458862, -0....
1	9742356831	[0.3207280337810516 4, 1.1068389415740967, 0.07...
2	B00005BPQ9	[-0.8029127717018127, 1.1512556076049805, 0.06...
3	B00006BN4U	[0.32304733991622925, 1.552841067314148, -0.17...
4	B00006FMLY	[-1.3537788391113281, 0.6959276795387268, -0.4...
...	...	...
6181	B01H3VFR6U	[-0.8585600852966309, 1.2054538726806 64, -0.49...
6182	B01H4G38JM	[-1.002007246017456, 0.6213456392288208, -1.11...
6183	B01H6IQING	[0.38772085309028625, 1.314347267150879, 0.187...
6184	B01H6JO1EI	[-0.7649469971656799, 0.6929209232330322, -0.9...
6185	B01HCV88MQ	[0.0764944776892662, 1.0583688020706177, -0.63...

[그림 1] Description 임베딩 값

4) 추천시스템 모델

본 논문에서는 총 3가지(BiVAE, NCF, Light FM)의 추천 시스템 모델을 이용하였다.

BiVAE의 경우 기존 VAE의 단점을 보완하기 위해 제안된 모델이다. VAE는 preference 데이터와의 two-way nature 불일치 때문에 부가 정보를 표현하기 어려움을 지니고 있다. 하지만 BiVAE의 경우 User-item Interactions의 Generative 모델과 다층 신경망을 사용하여 Parameterized된 한 쌍의 Inference 모델로 구성된다. 그렇기에 Dyadic 데이터의 양쪽에서 불확실성을 포착할 수 있어 기존의 단측 변형 자동 인코더에 비해 희소 선호도 데이터에 대한 견고성과 성능을 향상시킨다.

NCF(Neural Collaborative Filtering)는 기존의 협업 필터링 모델에 딥러닝을 추가한 모델이며 총 4개의 layer (Input Layer, Embedding Layer, Neural CF Layer, Output Layer)로 구성되어 있다. 잠재 요소(latent factor)간의 내적 연산을 대체하였고, 딥 러닝의 특징 중 하나인 비선형성을 사용해 유저와 아이템의 상호작용을 학습했다. 또한 layer를 쌓을수록 좋은 성능을 보인다.

LightFM은 Hybrid Matrix Factorization 모델로 Content-based 모델과 Collaborative Filtering 모델의 장점을 모두 가지고 있다. Collaborative 데이터와 User/Item 피처(Feature)를 학습 데이터로 사용한다. 유저와 아이템 각각의 피처와 둘 간의 상호작용을 고려할 수 있으며 Cold Start 문제를 완화시킬 수 있다.

5) 성능

Meta Data를 이용한 모델이 이용하지 않은 모델보다 더 높은 성능을 보였다. 그러나 NLP task를 거친 Meta Data 활용 모델의 성능보다 BiVAE 모델의 성능이 월등히 높게 나온 것을 확인할 수 있었다.

3. 결 론

확실히 Meta Data를 이용한 모델이 Meta Data를 이용하지 않은 모델보다 더 높은 성능을 보임을 확인할 수 있어, 기존의 목표는 달성하였다. 하지만 예상했던 것보다는 더 낮은 성능을 보여주었는데, 그 이유는 딥 러닝 모델에서 GPU의 한계로 데이터셋을 불가피하게 줄였기 때문이라 추측된다. 따라서 기존의 데이터셋을 모두 이용할 수 있었더라면 더 좋은 성능을 보여주었을 것이다.

이에 더해 BiVAE가 월등한 성능을 보인 이유를 예측해보자면 첫째로는 이미 언급했듯이 딥 러닝을 이용한 모델에서 자원의 한계로 데이터 수를 축소하여 더 높은 성능을 보여줄 수 없었다는 점이 있다. 실제로 모든 데이터셋을 이용한 딥 러닝 모델이 더 좋은 성능을 보일 가능성을 배제할 수 없다. 두 번째로는 BiVAE가 상당히 최신에 발표된 논문으로 기존 VAE의 단점을 보완하여 희소 선호도 데이터에 대해 견고성과 성능을 향상시킨 모델이기 때문에 다른 모델보다 우수하였을 가능성이 있다. BiVAE가 높은 성능을 보인 이유를 얻기 위해서 동일한 데이터를 이용해 기존 VAE 모델과 비교한 후속 연구를 진행할 예정이다.

참고 문헌

[1] Geng, Shijie & Liu, Shuchang & Fu, Zuohui & Ge, Yingqiang & Zhang, Yongfeng. (2022). Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)

[2] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

[3] Quoc-Tuan Truong, Aghiles Salah, Aghiles Salah. 2021. Bilateral Variational Autoencoder for Collaborative Filtering. In *WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining,* pages 292–300

[4] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, Tat-Seng Chua. 2017. Neural Collaborative Filtering. In WWW '17: Proceedings of the 26th International Conference on World Wide Web, pages 173–182

[5] Maciej Kula. 2015. Metadata Embeddings for User and Item Cold-start Recommendations