딥러닝 기반 영화 흥행 예측 및 영화 추천 모바일 시스템 개발

김경석, 장재준, 강현규[1)]
건국대학교 컴퓨터공학과
zkadhs12@naver.com, gihojpkl@naver.com, hkkang@kku.ac.kr

A mobile system development which has function of movie success prediction and recommendation based on deep learning

Kyeong-Seok Kim, Jae-Jun Jang, Hyun-Kyu Kang*
Department of Computer Engineering, Konkuk University

요 약

본 논문은 공공 데이터 Open API와 TMDB(The Movie Database) API를 이용하여 사용자의 선호 영화를 Google에서 제공해주는 Tensoflow로 인공신경망 딥러닝 학습하여 사용자가 선호하는 영화를 맞춤 추천하는 애플리케이션의 설계 및 구현에 대하여 서술한다. 본 애플리케이션은 사용자가 쉽게 영화를 추천받을 수 있도록 만들어진 애플리케이션으로 기존의 필터링 방식으로 추천하는 방식의 애플리케이션들과 달리 사용자의 취향을 딥러닝 학습을 통해 최적의 영화 Contents를 추천함과 아울러 기존 영화의 특성을 학습하여 흥행할 신규 영화를 예측하는 기능 또한 제공한다. 본 애플리케이션에 사용된 신규 영화 흥행 예측 모델은 약 85%의 정확도를 보이며 사용자 맞춤추천의 경우 기존 장르 추천이나 협업 필터링 추천보다 딥러닝을 통한 장르, 감독, 배우 등의 보다 세밀한 학습 추천이 가능하다.

주제어: 추천, 학습형 애플리케이션, 영화, 딥러닝

1. 서론

Google 딥마인드의 "AlphaGo" 이래로 machine learning이 많은 주목을 받고 있다. 이러한 인기에 힘입어 전 세계적으로 machine learning과 접목하려는 시도가 여러 분야에서 동시다발적으로 보인다[1 ~ 4].

본 연구는 여러 분야 중에서 영화 분야에 초점을 맞춰 딥러닝 알고리즘을 통한 사용자 맞춤 영화 추천 시스템의 개발에 목적을 둔다. 영화의 자체적인 특성을 학습에 활용하여 흥행하는 영화의 특징을 발견하고, 이를 토대로 개봉 예정인 영화에 대해서 흥행의 여부를 예측하고 사용자에게 추천하고자 한다(영화의 자체적인 특성은 영화의 외적인 특성을 제외하고 장르. 감독, 배우[5] 등의 내적인 요소를 의미함). 또한 시스템을 이용하는 사용자의 취향 데이터를 수집하고 이를 학습하여 사용자의 취향에 맞춘 영화를 추천하여 기존의 영화뿐만 아니라 신규 영화까지 폭넓은 추천을 할 수 있게 한다.

2. 연구 배경 및 방법

2.1 연구 배경

본 연구에서는 딥러닝과 접목할 분야로 영화시장을 선택했다. 영화 시장에서 본 연구와 마찬가지로 추천 시스템을 도입한 경우는 다수 존재하지만, 딥러닝보다는 협업 필터링을 이용하거나 신규 영화에 대해서는 간단한 설문조사에 따른 결과만을 제공하는 것이 대부분이다. 본 연구는 이러한 시도에서 더욱 나아가 사용자 측면의 데이터와 영화 자체적인 특성의 데이터를 딥러닝을 통해 학습하여 사용자 개인의 취향에 맞춰진 영화 추천과 신규 영화에 대한 추천을 제공하고자 한다.

2.2 연구 목적

본 연구의 일차적 목표는 영화의 특성에 따라 영화의 흥행을 예측하는 데 있다. 영화 자체의 속성에 따른 데이터를 바탕으로 딥러닝에 기반을 둔 예측 알고리즘 모델의 학습을 통해 새로 개봉할 영화들에 대하여 사전에 성공할 영화를 예측하고 추천하는 모바일 애플리케이션을 개발한다.

이차적 목표는 영화의 특성과 사용자의 선호도 데이터를 바탕으로 사용자에게 맞춤 추천을 함에 있다.

최종 목표는 딥러닝을 이용하여 기존 앱에서 제공하는

1) Corresponding Author

추천 기능을 딥러닝 알고리즘으로 보완하여 성능을 높이는 데 있다.

2.3 연구 방법

연구는 데이터적인 측면과 딥러닝 모델적인 측면에서 병행하여 진행한다. 데이터의 분석, 수집, 전처리, 기준 선정 과정을 통하여 학습 데이터를 확보함과 동시에 데이터에 적합한 구조적인 딥러닝 모델을 구성하고 데이터와 모델을 접목한다. 이후 실험적 테스트 과정을 거쳐 산출된 결과물의 질적 향상을 위하여 데이터와 모델 양쪽 측면에서 결과 분석 및 개선 방안을 논의하고 이를 다시 적용한다.

3. 데이터 수집 및 적용

3.1 데이터 수집

데이터의 수집은 한국 영상자료원 Open API[8], Kaggle Data set[9], TMDB[10]을 통해 일차적으로 진행하였고, crawling 기법을 활용하여 추가적인 데이터 수집 및 학습에 활용하기에 부적합한 데이터를 보완하는 방식으로 진행하였다. 이후 수집된 데이터의 분석 및 전처리 과정을 거쳐 학습 데이터로써 가공한다.

사용자의 취향 데이터의 경우, 사용자의 선호에 따라 수집되어야 하므로 사용자가 시스템을 사용하면서 선택한 "좋아요"와 "싫어요"를 기반으로 사용자 취향 데이터를 정의하고 해당 영화를 선호(1)와 비선호(0)로 구분하여 수집한다.

3.2 데이터 현황

Kaggle[9]과 TMDB[10]에서 승인받은 API를 통하여 전체적으로 44만 개의 open data set을 수집하였고, 학습에 활용할 수 있는 데이터를 분류하기 위한 필터링 작업과 보완 및 전처리 작업을 거쳐 전체적으로 17,000개의 data set을 확보하였다. 확보한 data set 중 80%(13,600개)는 학습에 활용하였고 나머지 20%(3,400개)는 학습 결과의 테스트 작업의 목적으로 활용한다.

3.3 학습 데이터 적용

영화의 흥행예측에는 영화 데이터의 속성 중에서 영화의 ID, 예산, 장르, 배우(주연 및 조연), 감독, 수익을 활용한다. 이 중 수익을 제외한 데이터가 모델의 feature로서 활용된다. 수익 데이터는 이들을 이분법적으로 흥행 및 비 흥행 영화로 구분할 수 있는 기준을 확립하여 label로서 활용된다. 기준에는 년도별 박스 오피스 100위 이내 및 제작비 대비 수익률을 사용한다. 순위는 데이터가 전 세계의 데이터인 점을 고려하여 박스오피스를 사용한다. 두 가지 기준 중의 하나 이상을 충족할 경우 흥행한 영화로 분류된다.

사용자 맞춤 추천에서는 사용자의 취향에 따라서 수집된 영화 데이터를 바탕으로 모델의 학습을 진행한다. 위와 마찬가지로 영화의 데이터 속성 중에서 영화의 ID, 예산, 장르, 배우(주연 및 조연), 감독을 활용한다. 결과 값은 선호와 비선호에 따라 1과 0으로 나타난다. 추천에는 사용자가 아직 경험하지 않은 영화들을 대상으로 적용하여 선호할 것으로 예측된 영화를 사용자에게 추천한다.

데이터 예시

Case	영화 아바타
식별 ID	19995(아바타)
예산	$237,000,000
장르	28(Action), 12(Adventure), 14(Fantasy), 878(Science Fiction)
배우	65731(Sam Worthington), 8691(Zoe Saldana), 10205(Sigourney Weaver),32747(Stephen Lang)
감독	2710 (James Cameron)
결과	1(흥행성공, 수익 기준)

4. 실험 및 분석

실험 및 분석은 딥러닝 모델[7]을 기본적인 로지스틱 회귀 모델부터 시작하여 weight초기화, dropout, ReLU, batch normalization의 기술을 점진적으로 적용하여 4가지의 모델을 만든 뒤 동일한 데이터로 5-Fold 교차검증 방법을 이용하여 정확도를 비교한다. 모델 검증은 데이터를 흥행 성공한 영화, 흥행 실패한 영화, 전체 영화 3부분으로 나누어 평균 정확도를 산출하여 비교한다. 또한 동일한 영화목록을 '좋아요'한 유저를 기준으로 기존 애플리케이션의 추천 알고리즘과 딥러닝을 적용한 추천 알고리즘의 결과를 비교한다.

4.1 1차 모델

1차 모델은 기본적인 로지스틱 회귀의 이항 분류 모델을 이용해 구현하였다. weight 값과 bias 값은 랜덤 값으로 초기화되었고 모델의 learning rate는 1% 총 100번 학습하였다.

흥행 성공 영화: [60%, 4%, 4%, 69%, 4%] || 평균 정확도 : 28%
흥행 실패 영화: [87%, 82%, 85%, 88%, 88%] || 평균 정확도 : 86%
전체 영화: [81%, 62%, 63%, 83%, 66%] || 평균 정확도 : 71%

그림 1. 1차 모델의 5-Fold 교차 검증 결과

그 결과 [그림 1]과 같이 흥행 성공 영화에 대해 평균 28%, 흥행 실패 영화에 대해 평균 86%, 전체 영화에 대해 평균 71%의 예측 정확도를 보였다.

4.2 2차 모델

2차 모델은 1차 모델에 입력층, 은닉층, 출력층이 각

1개씩인 심층 신경망을 적용하고 batch 시스템을 적용하여 학습하였다. 심층 신경망의 은닉층에서는 활성 함수로 ReLU 함수를 이용했으며 과적합을 방지하기 위해 dropout을 적용하여 70%씩 학습이 진행되도록 했다. 그리고 각 데이터의 분포를 고르게 하기 위해 표준화 전처리 작업을 했다. 모델의 learning rate는 0.5%이고 전체 17,000개의 데이터를 100개의 batch로 나누어 50 epoch 동안 학습을 진행했다.

흥행 성공 영화: [56%, 56%, 54%, 51%, 51%] || 평균 정확도 : 54%
흥행 실패 영화: [94%, 95%, 96%, 94%, 95%] || 평균 정확도 : 95%
전체 영화: [84%, 85%, 85%, 83%, 84%] || 평균 정확도 : 84%

그림 2. 2차 모델의 5-Fold 교차 검증 결과

그 결과 [그림 2]와 같이 흥행 성공 영화에 대해 평균 54%, 흥행 실패 영화에 대해 평균 95%, 전체 영화에 대해 평균 84%의 예측 정확도를 보였다.

4.3 3차 모델

3차 모델은 2차 모델에서 weight 값 초기화 방법으로 Xavier / He Initializer 적용했다. 또한 심층신경망의 은닉층을 기존의 1개에서 2개로 늘려 학습을 진행했다. 모델의 learning rate와 batch의 개수, 학습을 진행한 epoch는 2차 모델과 동일하다.

흥행 성공 영화: [64%, 63%, 57%, 62%, 63%] || 평균 정확도 : 62%
흥행 실패 영화: [93%, 92%, 94%, 92%, 93%] || 평균 정확도 : 93%
전체 영화: [86%, 84%, 84%, 84%, 85%] || 평균 정확도 : 85%

그림 3. 3차 모델의 5-Fold 교차 검증 결과

그 결과 [그림 3]과 같이 흥행 성공 영화에 대해서 평균 62%, 흥행 실패 영화에 대해 평균 93% 전체 영화에 대해 평균 85%의 예측 정확도를 보였다.

4.4 4차 모델

4차 모델은 3차 모델에서 배치 정규화를 추가로 적용했다. 심층 신경망의 각 은닉층에서 배치 정규화를 적용한 뒤 ReLu 활성화 함수를 적용하여 학습을 진행했다. 또한 배치 정규화에 dropout 효과가 기본적으로 내장이 되어 있기 때문에 기존에 적용된 은닉층의 dropout을 제거했다. 3차 모델과 동일하게 모델의 learning rate와 batch의 개수, 학습을 진행한 epoch는 2차 모델과 동일하다. 그리고 배치 사이즈가 100개로 매우 작기 때문에 배치 정규화를 진행하기 위한 momentum은 0.9로 높게 설정하였다[6].

흥행 성공 영화: [42%, 34%, 39%, 40%, 32%] || 평균 정확도 : 37%
흥행 실패 영화: [95%, 96%, 95%, 97%, 98%] || 평균 정확도 : 96%
전체 영화: [82%, 79%, 81%, 81%, 81%] || 평균 정확도 : 81%

그림 4. 4차 모델의 5-Fold 교차 검증 결과

그 결과 [그림 4]와 같이 흥행 성공 영화에 대해서 평균 37%, 흥행 실패 영화에 대해 평균 96%, 전체 영화에 대해 평균 81%의 예측 정확도를 보였다.

4.5 모델 비교분석

모델 / 평균정확도	1차	2차	3차	4차
흥행성공	28%	54%	62%	37%
흥행실패	86%	95%	93%	96%
전체	71%	84%	85%	81%

흥행 성공한 영화를 예측하는 평균 정확도는 1차 모델이 28%로 제일 낮았고 3차 모델이 62%로 제일 높은 모습을 보였다. 흥행 실패 영화에 비해 흥행 성공 영화의 예측 정확도가 낮은 이유는 흥행 실패 영화에 비해 흥행 성공 영화의 학습 데이터 수가 적기 때문이다. 전체적인 평균 정확도는 대체로 모델이 복잡해질수록 높은 정확도를 보인다. 하지만 4차 모델은 흥행 성공 영화에 대한 예측 정확도가 매우 낮은 모습을 보였다.

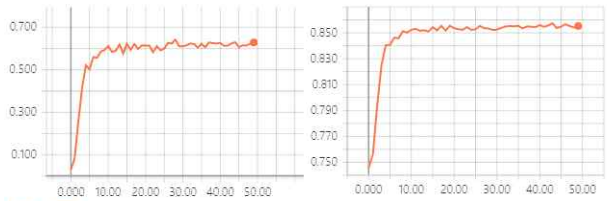

그림 5. 3차모델 정확도 좌:흥행성공, 우:학습 정확도

[그림 5], [그림 6]은 x축 epoch, y축 정확도로 이루어진 그래프이다. [그림 5] 에서 3차 모델의 경우 모델이 학습하면서 학습 정확도가 올라감에 따라 흥행 성공 검증데이터의 예측 정확도 또한 비례하면서 올라가는 모습을 볼 수 있다.

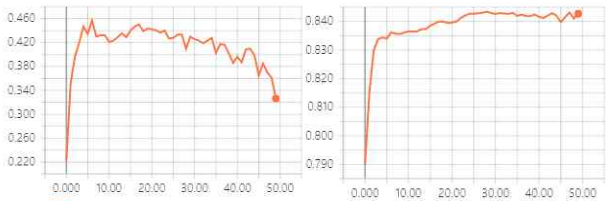

그림 6. 4차모델 정확도 좌:흥행성공, 우:학습 정확도

반면 [그림 6]에서는 4차 모델의 경우 모델이 학습하면서 학습 정확도가 올라가지만, 흥행 성공 검증데이터의 예측 정확도는 반비례하면서 내려가는 모습을 볼 수 있다. 따라서 현재 가지고 있는 데이터에 대해서는 3차

모델이 더 적합하다고 볼 수 있다.

4.6 사용자 맞춤추천 비교

title	seq
봄 오동 전투	17508
말모이	18029
완벽한 타인	17601
레슬러	17118
1987	16659
택시운전사	16150
공조	16047
베테랑	14173
극비수사	14737
그놈이다	15022
해적: 바다...	13930
소수의견	13864
감기	13825
미쓰 GO	13126
럭키	15711
NULL	NULL

그림 7. 좋아요 목록

['drama', 'action', 'comedy']

그림 8. 유저 선호도가 높은 장르

사용자가 [그림 7]과 같이 '유해진' 주연, 조연의 15개 영화를 '좋아요'를 눌렀을 경우 이 유저의 선호도가 높은 상위 장르 3개는 '드라마', '액션', '코미디'이다. 이 경우 기존 추천 알고리즘과 딥러닝을 적용한 사용자 맞춤 추천 알고리즘을 비교해보았다.

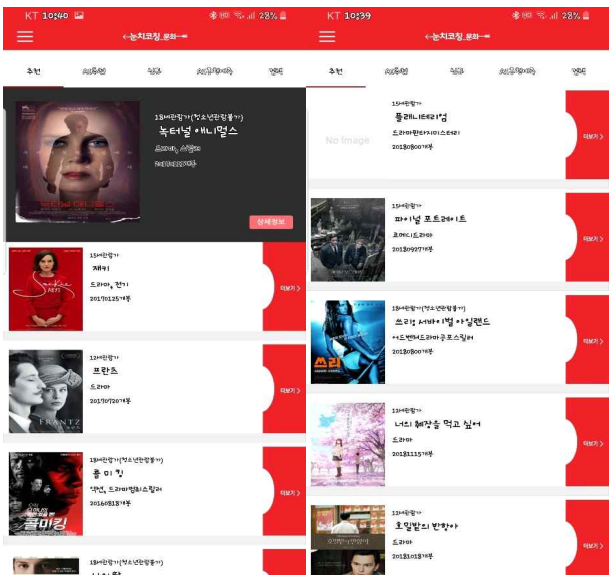

그림 9. 기존 추천 알고리즘 결과 화면

기존 추천 알고리즘의 경우 유저의 선호도가 가장 높은 '드라마' 장르 위주로 영화가 추천 되고 그 다음으로 '액션, '코미디' 등의 영화가 추천되는 것을 볼 수 있다. 하지만 '유해진' 주연, 조연의 영화는 추천 목록에 나타나지 않는다. 이는 기존 추천의 경우 추천에 장르만 적용될 뿐 그 외의 다른 감독이나 배우 정보는 반영되지 않기 때문이다.

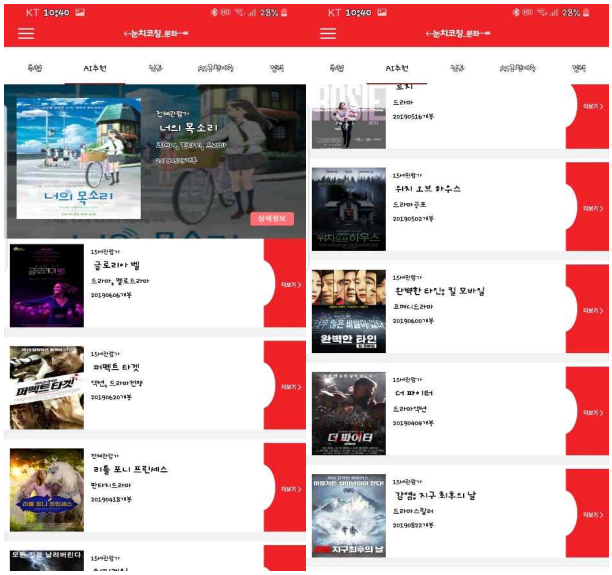

그림 10. 딥러닝 사용자 추천 결과 화면

반면에 딥러닝 알고리즘의 경우 기존 추천과 마찬가지로 유저 선호도가 높은 3개의 장르가 많이 추천되는 것뿐만 아니라 사용자가 '유해진'이란 배우를 선호하는 것을 학습하여 '완벽한 타인: 킬 모바일'이라는 유해진 주연의 영화를 추천하는 것을 볼 수 있다. 이는 유저의 선호도가 높은 장르뿐만 아니라 유저가 '좋아요'를 눌렀던 영화의 감독, 배우의 정보를 학습하여 영화를 추천했기 때문이다. 따라서 감독, 배우, 장르를 학습한 딥러닝 알고리즘 추천이 기존 장르만을 이용한 추천보다 사용자의 취향에 더 적합한 추천을 한다고 볼 수 있다.

5. 시스템 적용

5.1 눈치코칭_문화

"눈치코칭_문화"[11]는 영화, 뮤지컬, 연극, 콘서트, 국악, 무용, 미술, 문화의 날 정보 등 문화에 대한 전반적인 정보를 제공하는 모바일 기반의 추천 시스템이다. 영화, 공연에 대한 선호도를 외부 데이터베이스에 저장하고 이를 토대로 시스템의 자체적인 추천 알고리즘을 거쳐 사용자에게 추천한다. 해당 서비스는 Device에서 3G, WIFI 환경을 통하여 Open API, 외부 데이터베이스에 접근하고 정보를 받아옴으로써 수행된다.

5.2 영화 인공지능 추천 및 흥행예측

"인공지능_문화"는 "눈치코칭_문화"[11]에서 인공지능 알고리즘을 사용하여 추천 시스템을 조금 더 강화한 version이다. 기존의 모든 기능은 그대로 유지하면서 인공지능을 기반으로한 추천 및 흥행예측의 기능을 추가적으로 적용하여 사용자 맞춤 추천 서비스를 강화하였다.

구체적으로 "좋아요" 및 "싫어요"를 통한 사용자 선호 데이터 수집 기능 및 'AI 추천' 및 'AI 흥행예측' 서비스가 추가되었다.

5.2.1 Main activity

Main activity에서는 영화 및 공연 추천기능을 제공한다. 먼저 사용자 device의 고유 ID를 확인하고, 이를 통해 해당 사용자의 외부 데이터베이스 table에 접속하여 학습을 진행할 수 있을만한 취향 데이터가 충분히 수집되었는지 그리고 추천 영화 list가 완성되었는지 여부를 확인한다. 사용자 수집 데이터가 충분한 경우(15개 이상의'좋아요' 또는 '싫어요' 데이터) 인공지능 모델에서 1순위로 추천하는 영화가 가장 상단에 표시되고 2순위로 나온 영화가 다음에 위치한다. [그림 11]이 해당 경우에 해당한다. 그 다음 layout에는 기존에 기능에 따른 공연의 추천이 뒤를 잇는다. 처음 사용하는 사용자 혹은 학습에 필요한 데이터가 부족한 사용자의 경우에는 기존의 기능을 통해 추천한다. 인공지능 추천 및 일반 알고리즘 추천 여부는 1순위 추천 layout의 '인공지능 추천 Best' 또는 '추천 Best' mark를 통해 구분된다.

그림 11. AI 추천

그림 12. 기존 추천

5.2.2 영화 - AI 추천 탭

Movie activity의 AI 추천 탭은 main activity와 마찬가지로 먼저 사용자의 고유 ID를 통해 사용자의 취향 데이터의 충분한 수집 여부를 체크한다. 만약 데이터가 충분하다면, 외부 데이터베이스에서 추천 영화 list를 얻어오고(이는 main activity에서 제공하는 인공지능 추천을 포함한 전체 결과) 이를 바탕으로 한국 영상 자료원[8]의 API를 활용하여 추천하는 영화의 data를 얻어온다. 해당 data를 통해 생성한 list를 선호할 확률에 따라 정렬 하여 device 화면에 출력한다. 데이터가 부족할 경우, 사용자에게 해당 사실을 공지하는 toast message를 출력한 후 기존의 추천 기능을 통해 영화 list를 출력한다.

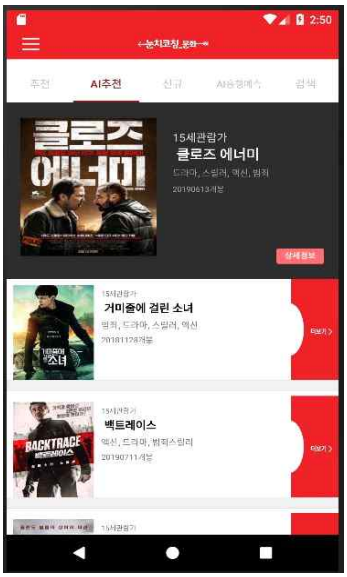

그림 13. AI 추천

5.2.3 영화 - AI 흥행예측 탭

Movie activity의 AI 흥행 예측 탭의 경우, 외부 데이터베이스에 저장된 개봉 예정인 영화에 대한 흥행 예측 결과 list를 가져오고 이를 흥행 확률에 따라 4가지 색상으로 구분하여 출력한다. (90% 이상 : Green, 89%~80% : Orange, 79%~70% : Yello, 69% 이하 : Red)
또한 영화의 이름 옆에 모델에서 예측한 해당 영화의 예상 흥행 확률을 함께 표시하여 사용자에게 제공한다.

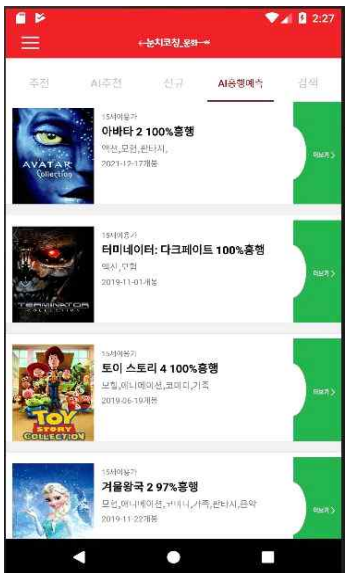

그림 14. AI 흥행예측

5.2.4 영화 - 상세보기

상세보기 activity는 list로 제공되는 영화에 대해서 상세한 정보를 제공하기 위한 기능이다. 영화명, 감독, 배우, 개봉일, 개봉 국가, 줄거리, 관람 등급, stills

cut, 포스터, runtime을 제공한다. 또한 "좋아요", "싫어요"를 통하여 사용자의 취향 데이터를 수집하는 역할을 수행한다. “좋아요”와 “싫어요”는사용자가 두 가지 버튼을 중복으로 선택할 수 없도록 설계되었으며 중복으로 선택하려고 할 경우 toast message로 해당 사실을 사용자에게 공지한다.

그림 15. 상세보기

6. 결론

신규 영화의 흥행 예측을 하기 위해 모델을 4차례에 걸쳐 개발하였다. 아주 간단한 1차 모델의 경우 학습이 제대로 되지 않는 모습을 보여주었고 가장 복잡한 4차 모델의 경우 현재 가지고 있는 데이터에 과적합 학습을 하여 적합하지 않다고 판단했다. 따라서 흥행 예측을 하는 모델은 3차 모델을 적용하는 것이 가장 적합한 것을 보여준다.

사용자 맞춤 추천의 경우 기존 알고리즘은 사용자가 선호하는 장르만 추천에 영향을 미칠 뿐 그 외 다른 정보는 추천에 반영되지 않는 모습을 볼 수 있었다. 하지만 딥러닝을 이용한 추천의 경우 사용자가 선호하는 장르뿐만 아니라 사용자가 선호하는 배우가 반영되는 것을 볼 수 있었다. 따라서 딥러닝을 이용하는 추천 방식이 좀 더 세밀한 추천을 할 수 있다.

딥러닝을 이용하여 학습하면 좀 더 세밀한 추천 효과를 볼 수 있다는 것을 알게 되었다. 3차 딥러닝 모델의 경우 흥행 성공 영화 예측에 대해서는 62%, 흥행 실패 영화 예측에 대해서는 93%, 평균적으로는 약 85%의 예측 정확도를 보인다. 향후 흥행 성공 영화 예측에 대해 약 80%의 정확도를 보이는 모델을 개발하여 전체적인 정확도를 90% 수준까지 올리는 것이 목표이다.

참고문헌

[1] 쿠지라 히코우즈쿠에, "머신러닝, 딥러닝 실전 개발 입문" , 위키북스, 2017.
[2] 김승현, 정용주, "처음 배우는 머신러닝 (사이킷런으로 기초부터 모델링 실전 예제 문제 해결까지)", 한빛미디어, 2017.
[3] 루카 마사론, 알베르토 보세티, "실전활용! 텐서플로 딥러닝 프로젝트", 위키북스, 2018,
[4] 바라스 람순다르, 레자 자데, "한 권으로 끝내는 딥러닝 텐서플로", 한빛미디어, 2018.
[5] W. Timothy WallaceAlan SeigermanMorris B. Holbrook, "The role of actors and actress in the success of films: how much is a movie star worth?", Journal of Cultural Economics, 10, pp.1-27, June 1993
[6] christian Szegedy, "BatchNormalization: Accelerating Deep Network training by Reducing Internal Covariate Shift" , https://shuuki4.wordpress.com/2016/01/13/Sergey Ioffe, arXiv:1502.03167, 2015.
[7] 김성훈, 모두를 위한 머신러닝/딥러닝 강의, http://hunkim.github.io/ml/
[8] 한국 영상자료원, KMDB Open API, https://www.kmdb.or.kr/
[9] Kaggle, https://www.kaggle.com/
[10] TMDB, https://www.themoviedb.org/
[11] 전재환, 이대영, 강현규, "기계 학습형 사용자 맞춤추천 앱 ‘눈치 코칭_문화’ 개발", 제29회 한글 및 한국어 정보처리 학술대회 논문집, pp.242-247, 2017.