|  |
| --- |
| **데이터사이언스개론**  데이터 사이언스의 적용 사례 REPORT |
|  |
| **컴퓨터공학과 201811259 배수빈** |

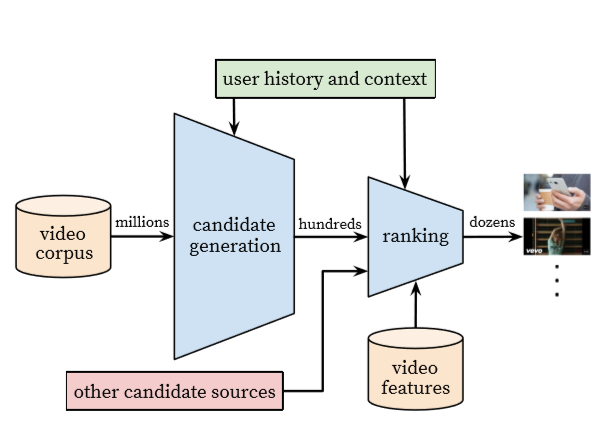
* **주제**

유튜브의 추천 시스템

* **주제 선정 이유**

코로나19로 사회적인 활동이 줄어듦에 따라 일상생활 속에서 유튜브 영상을 보는 일이 늘어나고, 내가 유튜브를 볼 때 원하는 영상을 검색해서 보는 경우보다 유튜브에서 추천하는 영상을 보는 경우가 많다는 것을 느낀 뒤 해당 알고리즘이 어떻게 개발 된 것인지, 어떤 원리인지 궁금해져서 선정하게 되었다. 실제로, 유튜브의 담당자의 인터뷰에 의하면 유튜브 이용자들의 시청시간 70%가 추천 기능에 의한 결과라고 한다. 따라서 해당 기능이 유튜브의 핵심적인 기술이라고 생각하고, 이의 개발과정과 원리를 알아보도록 하겠다.

* **유튜브의 추천 시스템**



* 구글 딥 러닝 인공지능 Brain과 심층 신경망(DNN) 모델 사용
* 엄청난 다중 분류 문제: 수많은 비디오 중 사용자가 오랜 시간 시청할 비디오 몇 개를 골라 내기
* 두 단계의 정보 검색 방법을 따라 2개의 심층 신경망으로 구성된 시스템

1. 후보 생성 네트워크 ( 사용자가 다음에 시청할 비디오 예측 )

2. 순위 평가 네트워크 ( 사용자의 비디오 시청시간 예측 )

* **1. 후보 생성 네트워크**

**1. 후보 생성 모델이 풀어야 할 문제 :**

사용자가 시청할 다음 비디오를 예측하는 것이다.

**2. 후보 생성 모델이 하는 일 :**

시청한 비디오 ID, 검색어, 사용자 정보 (위치, 나이, 성별 등) 등 유저의 활동 내역과 유저가 비디오를 시청한 문맥을 배우는 것이다. 모든 유저 정보는 벡터로 “embedding” (숫자화) 되어 심층 신경망에 입력되고 이 정보를 통해 모델은 수 만개의 비디오 사이에서 유저가 시청할 만한 몇 백개의 비디오를 분류한다. Softmax 다중 분류 함수를 활용해 분류된 비디오 각각의 시청확률 분석하여 시청 확률이 높은 비디오 선별한다.

**3. 후보 생성 모델 트레이닝 :**

딥 러닝 모델이 분류한 비디오가 사용자가 원하는 비디오인지 아닌 지 알려줌으로써 모델의 정확도 개선. 분류 성공: 분류한 비디오를 사용자가 끝까지 시청 / 분류 실패: 분류한 비디오를 사용자가 끝까지 시청하지 않음. 모델이 사용자가 끝까지 시청하는 비디오를 골라내도록 훈련.

**4. 후보 생성 모델이 새로운 비디오를 추천하게 하기 :**

사용자들은 자신과 관련 있는, 새로운 콘텐츠를 보고 싶어한다. 이를 위해 모델 트레이닝 시 업로드 된 지 얼마 안 된 비디오를 ‘분류 성공’ 카테고리에 포함시켜서 모델이 시간에 따른 비디오 인기도를 이해하게 한다. 이 방법을 통해 모델은 추천 후보 생성 시 업로드로부터의 시간을 반영한다. 새로운 비디오의 조회수를 올리거나 바이럴 콘텐츠를 퍼뜨리는 데 중요한 기술이다.

* **2. 순위 평가 네트워크**

**1. 순위 평가 모델이 풀어야 할 문제 :**

사용자가 추천 표시된 비디오를 얼마나 오래 시청할 것인가?

**2. 순위 평가 모델이 하는 일 :**

비디오와 유저에 대한 다양한 정보를 바탕으로 수백 개의 비디오 후보에게 점수 부여하고, 가장 높은 점수를 받은 비디오 몇 개를 사용자에게 보여준다.

각 사용자의 기준에 맞춰 사용자가 끝까지 시청할 확률이 가장 높은 비디오를 추천한다. 비교적 작은 숫자의 비디오만 분석하므로, 비디오에 대한 정보, 사용자와 비디오 간의 관계에 대한 다양한 정보를 처리할 수 있다.

**3. 다양한 데이터를 모델에 포함시키기 :**

모델에 포함시키는 데이터들은 다음과 같다. 연속형 데이터 (마지막으로 비디오를 시청한 때로부터의 시간, 이전에 비디오가 표시된 횟수), 이산형 데이터 (비디오 ID), Impression 데이터 (비디오에 대한 정보, 비디오가 점수를 부여 받을 때 마다 데이터 처리) , Query 데이터 (유저/문맥에 대한 정보. 입력이 있을 때마다 데이터 처리)

**4. 순위 평가 모델 트레이닝 :**

추천된 비디오를 보거나, 안 보는 2가지 경우의 수에 따라 비디오 분류한다. 비디오를 보는 경우 시청 시간에 따라 더 높은 점수를 준다. 유저가 표시된 비디오를 클릭하여 추천이 성공할 경우, 시청 시간에 따라 점수를 부여한다. 이와 같은 예로 모델을 트레이닝 해 사용자의 비디오 시청 시간을 근접하게 예측하는 함수를 생성하여 시청 시간이 높게 예측된 비디오를 추천한다.

**5. 두 모델이 공통적으로 보여주는 딥 러닝 인사이트:**

유투브의 추천 시스템 “Deep collaborative filtering model”. 다양한 정보를 연결해서 유저에게 가장 적합한 비디오를 추천한다. 심층 신경망을 겹겹이 쌓아 다양한 정보를 효과적으로 연결한다.

첫 번째 레이어에 다양한 정보를 입력하고 연결한다. 그 위에 수 겹의 히든 레이어 Rectiﬁed Linear Units (ReLU)를 연결한다. 레이어들이 연결되면서 직관적이지 않은 수 많은 데이터들 간의 관계가 연산된다. 데이터와 히든 레이어를 추가할수록 (레이어 depth가 증가할수록) 모델의 정확도가 증가한다. 방대한 데이터와 그 사이의 상호작용을 반영하여 시청 확률이 높은 비디오 추천한다.