내용 기반 유튜브 영상 추천 시스템 제안

윤영빈 윤준현 이태현 이현규 ° 박광훈
경희대학교 컴퓨터공학과
nogadahalf12@khu.ac.kr haring157@khu.ac.kr csws79@khu.ac.kr esot3ria@gmail.com
ghpark@khu.ac.kr

Proposal of Context-based Youtube Video Recommendation System

Yeongbeen Yun Junhyeon Yoon Taehyun Lee Hyungyu Lee
Department of Computer Science and Engineering, KyungHee University

요 약

내용 기반 유튜브 영상 추천 시스템(Context-based Youtube Video Recommendation System)은 기존 Youtube의 영상 추천 알고리즘을 개선하기 위한 시스템이다. 기존 알고리즘은 동영상 투고자가 직접 남긴 태그와 이용자의 과거 시청 데이터로만 의거하여 영상을 추천한다. 이에 본 논문에서는 기존 알고리즘에는 없었던 동영상 프레임 분석 기반의 자동 tag 추출 방법을 도입하여 보다 정확한 tagging을 하고, 이를 통해 이용자의 취향에 더 적합한 동영상 목록을 효율적으로 노출시킬 수 있는 시스템을 제안한다.

1. 서 론

'Youtube'(이하 유튜브)는 오늘날 인터넷 이용자들이 가장 애용하는 동영상 투고 및 시청 사이트이다. 유튜브를 통하여 동영상을 업로드하고 고정적으로 수입을 버는 'Youtube Creator'(이하 유튜버)라는 직업이 생길 정도이며, 이는 투고한 동영상의 조회 수와 유튜버 구독자 수가 높을수록 더 많은 광고 수익을 받는 시스템이다.

이러한 시스템으로 인하여 유튜브 본사도 자신들의 사이트 접속률을 높이기 위해 노력을 가한다. 그 활동 중 하나로 매년 'Youtube-8M Video Understanding Challenge'를 개최하는데, 이 대회는 동영상 dataset으로 tagging(이하 태그, 태깅)하는 연구와 개발이 주 목적이다. 우리는 이 대회의 목적을 본 논문의 주제로써 주목하고자 한다.

1년 동안에만 십 수억 개의 동영상이 업로드 되는 유튜브는 이용자의 관심을 가질만한 내용으로 영상을 추천해야할 필요가 있음에도 불구하고, 영상 내용 기반의 추천 알고리즘이 적용되지 않은 실정이다.

우리는 업로드 과정에서 동영상 프레임 중 일부를 추출하여 얻은 내용을, 기계학습을 이용한 유튜브 동영상 dataset 모델을 통해 그 동영상에 알맞은 태그를 달아주는 것이다. 그리고 태그들을 비교하는 작업인 word vector 기반의 자연어 처리로 최대한 관련된 동영상들을 밀집시켜, 이용자가 가장 흥미를 가질만한 동영상을 추천해주는 유튜브 영상 추천 시스템을 개발하고자 한다.

2. 기존 추천 알고리즘

한국언론진흥재단의 논문 '유튜브 추천 알고리즘과 저널리즘' [1] 에 따르면 유튜브의 추천 알고리즘에서 영상의 내용은 중요하지 않다고 한다. 현재 기술로는 내용까지 파악하기 어렵고, 사람이 직접 확인하기에는 시간이 너무 오래 걸리기 때문이다. 추천 영상 목록을 만드는 기준은 이용자가 과거에 시청했던 영상과 비슷한 주제, 그리고 사용자의 반응을 예측하여 만든다. 이용자가 보았던 동영상이나, '좋아요' 버튼을 클릭했던 동영상에 달려 있는 태그를 이용하여 같은 태그의 동영상을 추천해준다는 것이다.

유튜브 본사의 궁극적인 목적은 이용자를 유튜브에 최대한 오래 잔류시키는 것이다. 기존의 유튜브의 영상 추천 알고리즘은 이용자의 선호도를 기반으로 영상을 추천하고 있지만, 영상 자체의 내용은 분석에 활용하지 않는다. 그렇기 때문에 이용자의 성향에 정확하게 맞는 영상을 추천하기 힘들다.

이에 대해 우리는 기존 알고리즘보다 더 동영상 내용을 반영하여 태그를 달고 이용자에게 추천하는 것을 추구하고자 한다.

3. 내용 기반 추천 시스템

본 절에서는 내용 기반의 유튜브 영상 추천 시스템을 제안한다. 이 시스템은 현행 유튜브의 추천 알고리즘과는 다르게, 영상 내용을 분석하여 유사한 내용이 담긴 다른 영상들을 추천한다.

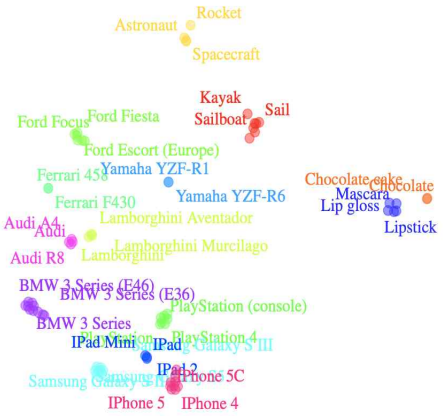

그림 1 Embedding된 word vector의 군집화

'Youtube-8M Video Understanding Challenge' [2] 등에서 사용된 Deep Learning 모델을 이용하면 영상을 분석하여 그에 맞는 태그를 붙일 수 있다. 최근 높은 성능을 보이는 모델의 경우 영상 분석 용도의 CNN과 자연어 처리 용도의 RNN을 함께 사용하는 Hybrid Architecture로 이루어진 경우가 많은데, RNN을 사용할 경우 각 tag에 해당하는 고정 크기의 word vector가 생성된다. [그림 1]에서는 의미적으로 연관이 있는 word vector들이 서로 가까운 공간에 embedding되는 현상을 보여 준다. 이 때 영상에 할당된 복수의 태그에 대한 word vector를 mean 연산 등 통계적 기법을 통해 하나로 합치면 영상을 대표하는 similarity vector인 video vector가 생성될 수 있다. 이렇게 복수의 태그를 합쳐 특정 영상을 대표하는 video vector를 생성한 뒤 영상과 함께 데이터베이스(이하 DB)에 저장해 두면, 각 영상의 video vector 간에 cosine similarity 등의 방법으로 서로 간의 수치적 유사도를 계산해낼 수 있다. 추후 이용자가 영상을 볼 때 데이터베이스(이하 DB)에 저장된 해당 영상의 video vector를 가져와, 가장 유사도 수치가 높은 상위 n개의 video vector를 찾을 수 있다. 이것이 가장 밀접한 내용을 가진 영상들의 목록이 되며, 내용상으로 가장 연관이 깊은 영상들을 이용자에게 추천할 수 있다. 새로운 영상이 업로드 되었을 때에도 해당 영상에 대한 video vector를 생성해 두면, 같은 방법으로 기존에 저장되어 있던 내용 상 밀접한 영상을 추천할 수 있다.

4. 시스템 구현 방안

해당 시스템은 누구나 쉽게 접근할 수 있고, 상호작용할 수 있도록 웹페이지를 통하여 구현하고자 한다. 편의를 위해 시스템을 Front-end와 Back-end로 나누어서 구현방안에 대해 설명할 것이다. 전체적인 시스템의 구성도는 다음 [그림 2]와 같다.

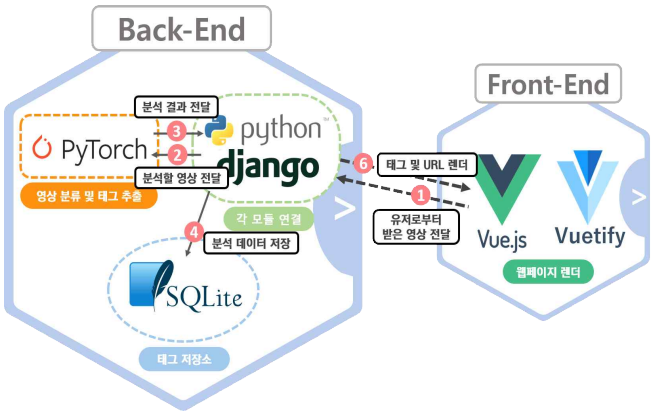

그림 2 웹페이지 시스템 구성도

웹페이지의 Front-end는 반응형 웹을 효과적으로 구현하기 위하여 Vue.js와 Vuetify를 사용하였고, Back-end는 PyTorch와의 호환성을 위하여 같은 python 프로그래밍 언어를 사용하는 웹 프레임워크인 django를 사용하여 구현하였다.

4.1. Front-end 구현

이 시스템의 Front-end는 서비스 이용자과 시스템이 상호작용하는 공간, 즉 웹페이지를 의미한다. 우리가 구현할 웹페이지는 [그림 3]에서 보이는 바와 같다.

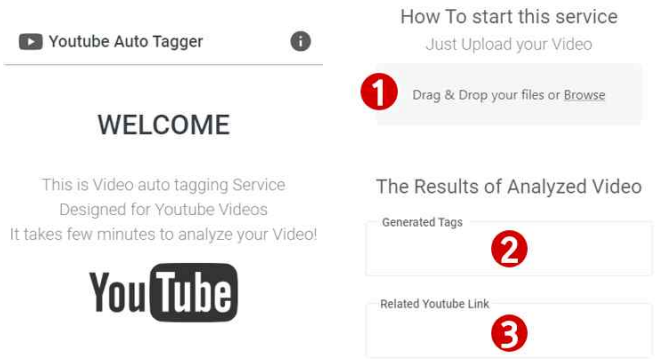

그림 3 웹페이지 UI

웹페이지에서의 이용자와 상호작용은 서비스를 태그 생성에 초점을 두기 위하여 간단하게 구성하였다. 서비스 이용자는 1번 박스를 클릭 또는 드래그 앤 드롭을 통하여 분석할 유튜브 영상을 시스템에 전달한다. 여기서 분석에 대한 자세한 내용은 4.2절에서 설명한다. Back-end의 시스템이 유튜브 영상을 성공적으로 분석을 완료했다면, 2번 박스에 해당 영상과 관련된 태그들이 그리고 3번 박스에는 해당 영상과 관련된 유튜브 링크들이 생성데이터베이스(이하 DB)된다.

4.2. Back-end 구현

본 시스템은 입력 영상과 유사한 영상 목록을 출력하기 위해 영상마다 하나의 video vector를 생성해 이용한다. 입력 영상은 시간에 따라 하나 이상의 장면을 보여준다고 전제한다.

영상을 학습시키기 위한 모델로는 Youtube-8M challenge에서 높은 성적을 보인 CCRL(Cross-Class Relevance Learning)을 사용한다.

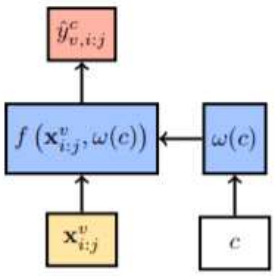

그림 4 CCRL Architecture

video vector를 생성하기 위해 다음의 과정을 거친다. 기존 영상분류 및 태깅 연구에서 사용되는 방식인 Multi-Stream으로 입력하기 위해 일정 시간 간격을 두고 영상의 프레임들을 추출한다. 전제가 되는 복수의 장면이 포함된 영상을 단일 장면만이 포함된 짧은 영상으로 분할해 분석하기 위해 전체 프레임들을 몇 개의 그룹으로 나눈다.

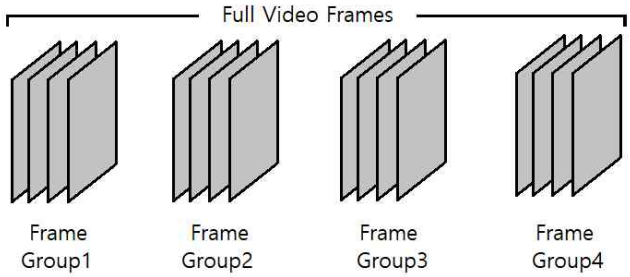

그림 5 Frame grouping

그림 6 Execute Label Weight

각 그룹을 학습 모델의 입력으로 사용하고, 그룹 당 하나의 label을 반환한다. 전체 label의 개수에서 각 label이 점유하는 비율을 계산하여 이를 영상에 존재하는 특정 label의 가중치로 사용한다.

각 label의 가중치는 'Word2Vec'을 통해 vector space에 존재하는 word vector에 곱하고, 같은 방법으로 얻어진 다른 label들의 결과와 합한다. 모든 결과를 합한 하나의 벡터를 입력 영상의 video vector로 사용한다.

video vector는 영상의 URL과 함께 DB에 저장된다. 저장된 video vector와 이후 사용자의 입력으로 새로운 영상이 입력되었을 때, 해당 영상의 video vector를 비교하여 유사도를 비교한다. 두 vector 간 유사도는 cosine similarity 등의 방법을 이용한다. 이때 데DB의 질의 속도를 향상시키기 위해 DB에 각 영상마다 가중치가 가장 높은 하나의 label을 추가적인 attribute인 Main Class로 저장한다.

Index	Video URL	Main Class	Video Vector

그림 7 Database Scheme

영상이 입력되면 word space에서, 입력 영상에서 가장 높은 가중치를 가진 label과 유사한 word들을 찾는다. 해당 word들을 조건으로 Main Class에 값이 일치하는 튜플을 검색하고, 검색 결과인 튜플들이 가진 video vector와 유사도를 비교한다. 일정 값 이상 유사도를 가지는 튜플은 추천하는 영상으로써 Video URL을 반환한다.

5. 결론 및 향후 연구

본 논문에서는 유튜브 동영상 업로드 과정 시의 영상 내용 기반 자동 tagging과 그 tag에 따른 영상을 이용자에게 추천해주는, 새로운 유튜브 영상 추천 시스템을 제안하였다.

이를 통해 Youtube Creator는 더 정확한 영상 tag와 높은 광고 효율을, Youtube 이용자는 더 취향에 맞는 관련 영상을, Youtube 본사는 더 높은 사이트 접속률을 기대할 수 있을 것이다.

향후 연구로는 Tagging의 정확도를 높이거나 전처리 과정을 단축시킬 수 있는 방안을 모색할 계획이다.

참 고 문 헌

[1] 오세욱, 송해엽, "유튜브 추천 알고리즘과 저널리즘", 한국언론진흥재단, 2019.

[2] Youtube-8M, "Youtube-8M Large-Scale Video Understanding Challenge", Google, 2019. https://research.google.com/youtube8m/workshop2019/

[3] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, Ilya Stanevich, "Cross-Class Relevance Learning for Temporal Concept Localization", 2019.