# 심층 시청각 복합 학습을 이용한 새로운 유튜브 조회수 예측 기법

박해연○ 이강훈 장연우 김형석 배성호
경희대학교 컴퓨터공학과
serahhy@khu.ac.kr, zxc8594@khu.ac.kr, 499510@khu.ac.kr, hsuper189@gmail.com, shbae@khu.ac.kr

# A New YouTube View Count Prediction Method using Deep Audio-Video Multimodal Learning

Haeyeon Park○ Ganghun Lee Yeonwoo Jang Hyeongseok Kim Sungho Bae
School of Computer Engineering, Kyunghee University

**요 약**

본 연구에서는 딥러닝 기반으로 유튜브(YouTube) 동영상을 분석하여 조회수를 예측하는 기법을 제안한다. 유튜브 동영상의 주제는 음식을 먹는 방송(이하 먹방)으로 한정하였고 총 1,076개의 동영상을 수집하였다. 본 연구에서 제안하는 딥러닝 모델은 영상 정보와 음성 정보의 융합 방법에 따라 3가지로 나뉘며, 각 모델의 최후 단에 구독자 수 및 게시 기간 정보를 함께 학습하여 동영상의 전반적인 조회수를 결정하는 외부 변인을 통제하였다. 영상 정보는 전이 학습을 이용하여, 음성 정보는 MFCC 변환을 이용하여 특징을 추출하였다. 학습 결과, 멀티미디어를 구성하는 영상 정보 및 음성 정보가 동시에 송출되는 특성을 가장 잘 반영한 intermediate level fusion 방법의 예측 성능이 가장 뛰어났다. 본 연구의 결과는 시청각 복합 학습을 통한 동영상 흥미도의 개괄적 예측이 가능함을 보인다.

## 1. 서 론

세계 최대 규모의 동영상[1] 플랫폼인 유튜브(YouTube)는 동영상 광고 시장에서 막강한 영향력을 미치고 있다. 유튜브의 국내 광고 수익은 2018년 상반기 기준 1,169억 원에 달하는 것으로 추정된다. 유튜브에서 발생되는 광고 수익은 파트너 프로그램을 통해 콘텐츠 제작자인 크리에이터(creator)와 분배하는 구조이며 동영상의 조회수가 증가할수록 크리에이터의 수익도 커진다. 이에 따라 크리에이터들은 보다 높은 조회수를 얻기 위해 더욱 좋은 품질의 영상 효과, 장비, 자막, 섬네일(Thumbnail) 제작 등의 노력을 기울인다. 그러나 국내 주요 크리에이터 3,000명이 2017년 업로드한 246,550개의 동영상 중 100만 이상의 조회수를 기록한 동영상은 5,955개로 전체 동영상의 약 2.42%에 불과하다. 따라서 본 연구에서는 크리에이터들이 동영상 업로드 전 사전 평가를 진행할 수 있도록 동영상 자체 정보를 중심으로 딥러닝 모델을 통해 분석하여 조회수를 예측하는 기법을 제안한다. 영상 정보와 음성 정보, 메타데이터를 이용하여 분석을 진행하였으며, 동영상의 범주에 따라 그 특성에 큰 차이가 있으므로 학습 동영상의 범주를 먹방으로 한정하였다.

## 2. 관련 연구

머신러닝 기반으로 동영상의 조회수를 예측하는 기법과 관련된 다양한 연구가 있다. 연구 [1]은 머신러닝 기반으로 유튜브 동영상이 가진 섬네일, 구독자 수, 제목 등의 메타데이터 분석을 통해 각 메타데이터가 조회수에 미치는 영향을 분석한다. 연구 [2]는 딥러닝 기반의 드라마 동영상 및 실시간 시청자 행동 분석을 통해 실시간 조회수를 예측하는 모델을 제시한다.

동영상은 영상 정보와 음성 정보가 동시에 송출되므로 유튜브 동영상 학습 시 두 정보를 적절히 종합하여 분석하는 정보 융합 기법이 요구된다. 연구 [2]에서 제시한 모델은 영상 정보와 음성 정보의 융합 단계를 기준으로 나누어 학습한다. 연구 [3]에서 제시한 모델은 영상, 음성 및 문자 정보를 종합하여 학습의 특정 구간에서 역전파의 적용 여부를 기준으로 나누어 학습한다. 연구 [4]에서 제시한 모델은 학습 시 4가지 센서 정보의 융합 단계를 기준으로 나누어 학습한다.

연구 [1]은 메타데이터만을 입력 정보로 하여 동영상의 조회수를 예측하였다. 하지만 연구 [1]은 동영상 게시 이후 하루가 지난 시점의 조회수를 입력 정보로 포함하므로 유튜브에 게시되지 않은 동영상에 대한 조회수는 예측할 수 없다. 연구 [2]에서는 드라마 동영상의 실시간 조회수를 영상 정보와 음성 정보로 나누어 예측하였다. 그러나 학습되는 동영상은 실시간 조회수를 포함하며 길이가 동일하므로, 단일 조회수를 가지며 길이가 다양한 일반적인 동영상에는 적용에 어려움이 있다.

본 연구에서는 유튜브 동영상의 영상 정보 및 음성 정보를 중심으로 소수의 메타데이터를 함께 이용하여 다양한 길이를 가지는 동영상의 조회수를 보다 정확하게 예측하는 모델을 제안한다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업의 연구결과로 수행되었음 (No. 2017-0-00093).

1) 본 연구에서는 동영상을 영상 및 음성 정보의 종합으로 정의한다.

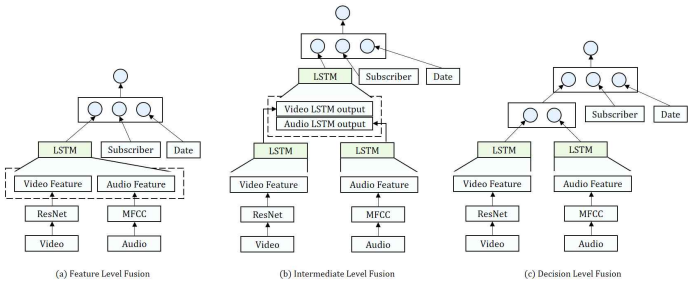


그림 1. 모델 구조

## 3. 제안 방법

### 3.1 데이터 수집

본 연구에서는 게시기간이 2주 이상이고 길이가 3분이상인 동영상을 학습 대상으로 선정하였으며, Google에서 제공하는 YouTube Data API를 통해 약 2,000개의 동영상을 수집하였다. 이후 모든 동영상을 검토하여 학습 목적에 적합하지 않은 동영상 924개를 제거하여 총 1,076개의 동영상을 선정하였다.

### 3.2 데이터 전처리

수집한 각 동영상은 대부분 초당 30개의 프레임을 가지며 전체 길이는 3분부터 1시간 30분까지 다양하다. 다양한 길이의 동영상을 학습하는 방법으로 각 동영상 정보에 패딩(padding)의 적용을 고려할 수 있다. 그러나 길이가 가장 짧은 동영상과 긴 동영상의 전체 프레임 수 차이는 약 156,600장으로 패딩을 적용할 경우 과도한 패딩으로 인한 학습 저하가 발생할 수 있다. 따라서 본 연구에서는 동영상을 특정 구간으로 잘라내어 모두 3분으로 통일하였고 초당 하나의 프레임과 MFCC (Mel Frequency Cepstral Coefficient)를 추출하였다. 또한, 메타데이터가 포함하는 정보 값은 성격에 따라 수의 범위가 다르므로 정규화를 적용하여 조정하였다.

#### 3.2.1 영상 데이터

모든 동영상에 대해 길이가 30분 미만인 동영상은 처음, 중간, 끝으로 구간을 나누어 각 1분을, 30분 이상인 동영상은 처음, 중간 1, 중간 2, 중간 3, 끝으로 구간을 나누어 각 36초를 취해 길이가 3분인 동영상을 생성하였고 잘라낸 동영상의 영상 정보에 대하여 초당 하나의 프레임을 추출하였다. 이후 추출한 각 프레임을 ImageNet 데이터셋에 대한 학습을 마친 ResNet34 모델의 마지막 fully-connected layer 직전 층까지 통과하여 얻은 $180\times2048$의 특징 벡터를 영상 입력 정보로 취하였다.

#### 3.2.2 음성 데이터

MFCC는 음성 처리에 널리 사용되며, 본 연구에서도 해당 알고리즘을 3.2.1에서 잘라낸 동영상으로부터 추출된 음성에 적용하였다. sampling windows의 크기는 25ms, 연속되는 windows 사이의 step은 10ms로 설정하여 초당 100개의 MFCC 벡터를 추출하였다. 이후 추출된 각 100개 MFCC 벡터의 평균을 구하여 얻은 $180\times13$의 음성 특징 벡터를 음성 입력 정보로 취하였다.

표 1. 모델 성능

| 모델 | MSE |
|---|---|
| Audio-LSTM | 6.55 |
| Video-LSTM | 5.75 |
| Feature-level fusion | 5.56 |
| Intermediate-level fusion | 4.97 |
| Decision-level fusion | 5.03 |

#### 3.2.3 메타 데이터

본 연구는 동영상 자체 정보를 중심으로 분석하여 조회수를 예측하는 것이 목표이다. 그러나 실제 유튜브 동영상의 조회수는 구독자수가 높을수록, 게시 기간이 길수록 높아진다. 따라서 동영상의 전반적인 조회수를 결정하는 구독자 수와 게시 기간을 학습 데이터에 포함하여 변인을 통제하였다. 구독자 수 및 게시 기간은 서로 정보의 성격이 달라 수의 범위에 차이가 있다. 이 차이는 학습 저하를 유발할 수 있어 두 정보에 MinMax Scale을 적용하여 메타데이터 입력 정보로 취하였다.

조회수 정보는 본 연구에서 제안하는 모델의 예측 목표이다. 수집한 동영상의 조회수는 지수적으로 증가하였으며 좌측으로 편향된 정규분포 형태를 보였다. 이는 조회수를 일정한 범위로 나누었을 때 범위대가 높아질수록 조회수 차이에 대한 민감도가 감소함을 의미한다. 이에 따라 발생할 수 있는 학습 시 편차를 최소화하고자 조회수에 자연 상수 $e$를 밑으로 하는 로그함수를 적용하여 정답 레이블로 취하였다.

### 3.3 모델 구조 및 학습

본 연구에서 제안하는 모델은 음성, 영상, 메타데이터(구독자 수, 게시 기간)의 3가지 성격의 입력 정보를 가지며, 음성 정보와 영상 정보를 융합하는 방식에 따라 그림 1과 같이 3가지 모델로 나누어진다. 모든 모델에서 LSTM 입력의 시퀀스 길이는 영상 정보의 프레임 수와 동일한 180이다. 또한, 모든 모델에서 최종 LSTM으로부터 출력된 마지막 시퀀스 정보와 메타데이터를 fully-connected layer에 통과시켜 출력된 정보를 융합하여 조회수를 예측한다. 그림 1.a의 FLF (Feature Level Fusion) 모델은 음성 및 영상 정보에서 추출한 특징 벡터를 융합한 정보를 최종 LSTM의 입력으로 한다. 그림 1.b의 ILF (Intermediate Level Fusion) 모델은 음성 정보와 영상 정보를 각각의 LSTM에 입력하여 출력된 정보를 시퀀스별로 융합한 후 최종 LSTM의 입력으로 한다. 그림 1.c의 DLF (Decision Level Fusion) 모델은 음성 정보와 영상 정보를 각각의 최종 LSTM에 입력하여 출력된 마지막 시퀀스 정보를 융합한다.

추가적으로, 영상 정보와 음성 정보의 융합이 예측 성능에 미치는 영향을 파악하고자 음성 정보 또는 영상 정보만을 LSTM에 통과시킨 후 메타데이터와 융합하여 조회수를 예측하는 Audio-LSTM, Video-LSTM 모델을 설계하였다.

설계한 모델은 단일 조회수 예측이라는 회귀 문제 해결을 목표로 한다. 이에 따라 모든 모델에서 목적함수로

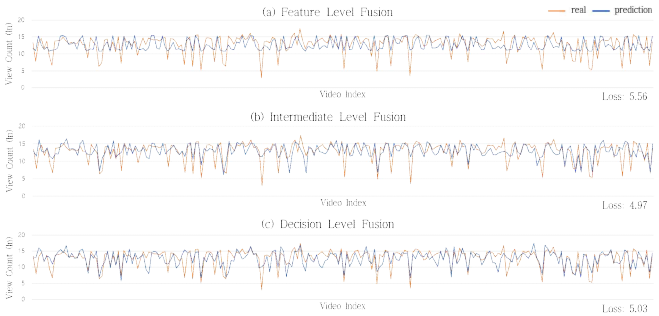


그림 2. 융합 모델의 학습 적합도

MSE를 채택하였다. 전체 데이터 중 850개는 학습 데이터이며, 226개는 시험 데이터이다. 모델 내부 뉴런 층, 미니배치 크기, 에폭 수 등의 하이퍼파라미터는 모델마다 그 값을 다양하게 조정하여 최적화하였다.

## 4. 학습 결과 및 분석

표 1은 각 모델의 성능을 측정한 결과이다. 3가지 융합 모델 모두 Audio-LSTM, Video-LSTM 모델보다 예측 성능이 좋았다. 이는 동영상 학습 시 음성 정보와 영상 정보를 융합하는 것이 단일 정보만을 이용하는 것보다 효과적임을 보인다. FLF 모델의 오차는 5.56으로 융합 모델 중 예측 성능이 가장 좋지 않았다. 이어 ILF 모델과 DLF 모델의 오차는 각각 4.97, 5.03으로, ILF 모델의 예측 성능이 가장 뛰어났으며 DLF 모델이 뒤를 이었다.

그림 2는 제안된 3가지 모델의 예측 적합도를 나타낸다. FLF 모델은 조회수의 예측 범위가 다른 모델에 비해 좁고 일정하였다. FLF 모델의 좋지 않은 예측 성능은 서로 다른 특징 공간을 가지고 있는 정보의 직접적인 융합이 학습의 한계로 작용할 수 있음을 보인다. ILF 모델은 영상 정보와 음성 정보가 시퀀스 단위로 융합되었기에, 동영상 시청 시 시각정보와 청각정보를 매 순간 통합하여 받아들이는 사람의 인지 과정을 가장 잘 반영하는 모델이다. ILF 모델의 뛰어난 예측 성능은 사람의 시청각 정보를 수용하는 방식을 잘 반영한 모델이 동영상 호감도의 지표인 조회수를 예측하는 데에 효과적임을 보인다. DLF 모델은 ILF 모델보다는 오차가 간소하게 높았으나 가장 동적인 예측을 보였다. DLF 모델의 높은 예측 성능은 각 영역을 표현하는 정보를 전체적으로 융합하여 학습할 수 있음을 보인다.

그림 3은 예측 결과 예시를 나타낸다. 그림 3.c와 같이 예측이 부정확한 동영상들은 다른 학습 데이터와 다소 다른 특징을 보였다. 예를 들어, ASMR[2]과 같이 영상보다는 음성에 치중된 동영상이거나, 음식을 먹는 모습보다는 코믹한 연출에 초점을 맞춘 동영상인 경우, 또는 사회적으로 큰 화제가 되었던 음식과 관련한 동영상인 것을 확인할 수 있었다. 위 동영상들의 조회수는 동영상 자체 정보보다는 인간의 내재적 감각이나 사회적 현상과 관련이 깊다. 또한, 외국인의 먹방 동영상에 대해서도 예측 능력이 떨어지는 현상은 국가마다 유튜브 사용자들의


(a) 조회수가 높은 동영상에 대한 예측 정확도가 높은 결과

(b) 조회수가 낮은 동영상에 대한 예측 정확도가 높은 결과

(c) 예측 정확도가 낮은 결과

그림 3. 예측 결과 예시

이용 행태 및 조회수를 이끌어 내는 동영상의 요소가 다르며 이를 반영할 수 있는 다양한 국가의 먹방 동영상 정보의 부족에 기인한 것으로 보인다.

## 5. 결론 및 향후 계획

제시된 학습 결과는 동영상 자체 정보를 통해 동영상의 조회수에 대한 개괄적 예측이 가능하다는 것을 보인다. 또한, 동영상 정보 학습 시 정보의 융합 방법에 따라 모델의 예측 성능이 달라질 수 있으며 사람의 영상 및 음성 정보 인지과정을 잘 반영해야 함을 보인다.

향후 최적 조회수 예측 모델의 가중치 값을 살펴 각각의 입력 데이터들이 조회수에 어떤 영향을 주는지 밝혀내고자 한다.

## [참고 문헌]

[1] William Hoiles, "Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 7, 2017.

[2] Xinpeng Chen, "Fine-grained Video Attractiveness Prediction Using Multimodal Deep Learning on a Large Real-world Dataset", Wuhan University, Tencent AI Lab, National University of Singapore, University of Rochester, 2018.

[3] Onno Kampman, "Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction", Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, 2018.

[4] Jing L, "An Adaptive Multi-Sensor Data Fusion Method Based on Deep Convolutional Neural Networks for Fault Diagnosis of Planetary Gearbox.", School of Mechanical Engineering, Tianjin University, Tianjin, 2017.

2) Autonomous Sensory Meridian Response, 주로 청각을 중심으로 하는 시각적, 청각적, 촉각적, 후각적, 혹은 인지적 자극에 반응하여 나타나는, 심리적 안정감이나 쾌감 따위의 감각적 경험을 일컫는 말.