JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 6, pp. 1123-1130, Jun. 2022

Check for updates

유튜브 알고리즘 요인 탐색을 위한 역공학설계 연구 – 머신러닝과 딥러닝 응용을 중심으로

김 철 년[1] · 배 승 주[2] · 하 윤 수[3] · 이 상 호[4*]
[1]케이티 차장 [2]경성대학교 미디어콘텐츠학과 외래교수 [3]허브스코프 대표이사 [4*]경성대학교 미디어콘텐츠학과 교수

A Study on the Reverse Engineering Design for Exploring YouTube Algorithm Factors - Focusing on Machine Learning and Deep Learning Application

Chul-Nyuon Kim[1] · Seung-Ju Bae[2] · Youn-Soo Ha[3] · Sang-Ho Lee[4*]

[1]Deputy Director, KT Corporation, Seoul 05552, Korea
[2]Adjunct Professor, Department of Media Content, Kyungsung University, Busan 48434, Korea
[3]CEO, HubScope, Seoul 04785, Korea
[4*]Professor, Department of Media Content, Kyungsung University, Busan 48434, Korea

[요 약]

본 연구는 유튜브 알고리즘 요인 탐색을 위한 역공학설계 방식의 연구로서 머신러닝과 딥러닝 응용을 중심으로 진행하였다. YouTube는 콘텐츠 제작자에게 가장 인기 있는 콘텐츠 플랫폼이자 광고 플랫폼이다. 전세계의 수많은 유저들이 사업목적과 오락목적으로 유튜브를 이용함에도 불구하고, 유튜브의 알고리즘은 공개된 바가 없기에 연구자들은 이를 역공학 설계방식으로 탐색해보고자 하였다. 따라서 머신러닝 방식의 설계를 통해 동영상 콘텐츠의 조회수와 상관관계가 높은 변인들을 다수 확인하였다. 또한 이들 변인들을 이용한 머신러닝과 딥러닝 예측의 정확성을 확인하였다. 연구자들은 본 연구의 결과가 조회수와 구독자수를 늘리는 데 중요한 변인이 무엇인지에 대한 통찰력을 제공하는 등 실무적 측면에서 긍정적인 기여를 할 것으로 기대한다.

[Abstract]

This study deals with the reverse engineering design method for the search for factors in the YouTube algorithm, focusing on the application of machine learning and deep learning. YouTube is the most popular content platform and advertising platform for content creators. Despite the fact that many users around the world use YouTube for business and entertainment purposes, YouTube's algorithm has not been disclosed. Thus researchers tried to explore it using a reverse engineering design method. Therefore, many variables highly correlated with the number of views of video content were identified through the design of the machine learning method. Additionally, the prediction accuracy of machine learning and deep learning was checked by the variables. Researchers expect that the results of this study will make a positive contribution in practice, such as providing insight into what are important factors for increasing the number of views and subscribers.

색인어 : 유튜브, 딥러닝, 머신러닝, 역공학, 알고리즘

Keyword : YouTube, Deep Learning, Machine Learning, Reverse Engineering, Algorithm

http://dx.doi.org/10.9728/dcs.2022.23.6.1123

Received 23 April 2022; **Revised** 23 May 2022
Accepted 25 May 2022

***Corresponding Author; Sang-Ho Lee**

Tel: +82-51-663-5204
E-mail: leeshow@empas.com

Ⅰ. 서 론

본 연구는 유튜브 알고리즘 요인 탐색을 위한 역공학설계 방식의 연구로서 머신러닝과 딥러닝 응용을 중심으로 진행하였다. 모바일 통신환경의 저변이 확대되고 네트워크의 속도가 증가함에 따라 다양한 미디어가 생겨나고 있으며, 1인 미디어도 급속히 증가하고 있다. 접근과 사용이 용이하고 정보를 받아들이기 쉬운 형태로 제공하고 감각적으로 더 자극적인 멀티미디어 콘텐츠를 제공하는 미디어가 선호도가 높은데, 유튜브(YouTube) 플랫폼은 미디어 생산과 소비가 활발하게 일어나는 대표적인 미디어이다. 유튜브는 미디어 소비자가 가장 선호하는 멀티미디어 제공 플랫폼이자 검색 포털로 콘텐츠 크리에이터(creator)들이 수익을 창출하기 위해 활동하는 미디어 생태계이다.

유트브는 파트너 프로그램을 통해 콘텐츠 크리에이터들이 광고 수익, 채널 맴버십, 상품 섹션, Super Chat 및 Super Sticker, 유튜브 Premium 수익을 창출할 수 있는 기능을 제공하고 있다[1]. 유튜브는 자체적인 알고리즘을 통해 1인 크리에어터에게 제공할 수익모델을 만드는데 구독자 수, 영상의 순위, 2차 채널로의 확산, 댓글의 반응, 브랜드 쿼리 변동 등을 반영하여 수익을 계산한다[2]. 유트브의 수익창출 알고리즘은 공개된 바가 없으며 지속적으로 변경되고 있다. 현재 수익창출 방식은 기본적으로 콘텐츠의 조회수와 구독자 수가 증가하면 수익도 높아지는 구조로 되어 있다.

콘텐츠의 조회수를 높이기 위해서는 기본적으로 사용자가 유튜브에 접속했을 때 추천 영상이나, 검색을 했을 경우 상위에 노출이 되어야 사용자가 해당 영상을 시청할 가능성이 높아진다. 그러나 유튜브의 영상 추천이나 검색 상위 노출에 대한 알고리즘을 공개하지 않고 동영상은 제목, 설명, 콘텐츠의 검색어와의 연관성, 높은 상호작용 등 다양한 요소를 기준으로 순위가 매겨진다고 설명하고 있다[1]. 유튜브 영상 검색 시 영상의 정렬은 관련성, 업로드 날짜, 조회수, 평점의 4가지 기준으로 가능한데, 업로드 날짜, 조회수는 명확하게 확인이 가능하나 기본 정렬기준인 관련성과 평점은 그렇지 않다. 다만 알고리즘에 머신러닝, 딥러닝 기법을 활용한다는 것만이 알려져 있다.

본 연구에서는 역공학 설계방식으로 머신러닝을 활용하여 검색 시 관련성이 높은 변인들을 확인하고 머신러닝과 딥러닝을 통해 예측 정확도를 검증하고자 한다. 유튜브상의 영상물에는 콘텐츠뿐만 아니라 다양한 관련 정보들이 포함되어 있으며, API 연동을 통한 크롤링을 통해 영상에 대한 많은 정보들을 수집할 수 있다. 영상과 함께 있는 많은 정보들을 변인들로 하여 어떠한 정보들이 관련성이 높은지, 그리고 영상의 조회수 및 채널의 구독자수와 상관관계가 높은 변인들을 도출할 것이다.

유튜브에 조회수와 구독자수와 연관이 높은 변인들, 영상 검색 시 상위노출과 관련이 높은 변인들이 무엇인지 통찰력을 제공함으로써 조회수와 구독자를 늘리는 실무적인 측면에 기여할 것으로 기대한다.

Ⅱ. 선행연구

2-1 알고리즘 연구

최근 유튜브 알고리즘의 구조적 이해와 요인을 탐색하는 일부 연구에서 역공학설계 등 다양한 방식의 분석방법론들이 적용되고 있다. 유튜브의 OpenAPI로부터 영상 데이터의 수집 활용이 가능해짐에 따라 수집과 분석이 용이한 노드엑셀(NodeXL)을 이용한 연구[3],[4]와 YouTube Data Tools 프로그램을 이용한 연구[5],[6]를 통해 연구자들은 영상들의 관계 데이터를 활용하여 유튜브 알고리즘의 구조적 이해를 탐색하는 연구가 확인되었다. 또한 사용자의 유튜브 영상 재생 목록으로부터 도출한 소셜데이터에 기반하여 알고리즘을 추정하는 연구[7]와 유튜브 공간에서 검색에 따른 추천리스트의 상위 랭크된 영상들의 정렬 기준을 탐색하는 연구[8]등이 보고되었다. 그리고 유튜브에서 인기 영상을 구성하는 콘텐츠 요소들 간의 상관성을 분석하는 연구[2]의 경우에서는 구독 채널, 동영상 재생 목록, 업로드 동영상 목록 및 댓글(수), 키워드, 조회 수, 좋아요 수와 싫어요 수 등의 변인들을 이용하고 변인들 간의 상관성과 상대적 중요도 분석을 통해 알고리즘 요인을 파악하는 연구 등이 관측되었다. 그 외에도 유튜브의 알고리즘과 관련하여 큐레이션과 역공학 설계 방식의 연구를 수행한 수편의 연구가 보고되었다[9]-[11]. 본 연구는 생물학의 신경망을 수학적으로 모델링한 인공지능의 관점을 적용하였다. 즉, 머신러닝을 활용하여 변인들 간 상관관계를 확인하고, 머신러닝과 딥러닝 방식으로 예측 알고리즘을 구현하고자 한 것이다.

2-2 머신러닝

머신러닝(machine learning)은 생물학의 신경망 특히, 인간의 뇌 구조의 모방을 원천으로 컴퓨터가 학습을 통해 스스로 지식을 축적하도록 설계된 알고리즘이다. 머신러닝의 학습과정은 컴퓨터가 주어진 학습데이터로부터 추출된 특징(feature)을 학습(training/learning)하고 결정경계(decision boundary)를 찾아 이후 입력되는 새로운 데이터의 특징 값에 기반하여 데이터의 클래스를 결정(test/generalization)하는 순서로 진행된다. 머신러닝은 주어진 데이터의 클래스를 구분하는 패턴 인식(pattern classification)문제와 일련의 연속 값을 추정하는 회귀(regression)문제를 해결하는데 사용된다. 학습방법에 따라 지도학습(supervised learning), 비지도학습(unsupervised learning), 반지도학습(semi-supervised learning), 강화학습(reinforcement learning)의 4가지로 구분된다. 지도학습은 문제와 답을 모두 제공하는 방식으로 입력값에 대한 출력값을 알고 이에 대한 피드백을 통해 학습하는 방식이며, 분류(classification)모델

과 회귀(regression)모델에서 사용된다[12]. 분류모델은 학습데이터의 레이블(x, y) 중 하나가 결과값이며 알고리즘에 따라 기저벡터머신(support vector machine), 의사결정트리모델(decision tree model)등이 있다. 회귀모델은 학습데이터에서 도출된 함수식으로부터 계산된 임의의 값을 추정·예측하는 것에 사용된다.[13] 비지도학습은 입력값과 출력값을 제공하지 않고 데이터 패턴문제를 추정하며, 반지도학습은 입력값과 출력값의 데이터(labeled data)와 모르는 데이터(unlabeled data)를 함께 사용하는 학습방식이다. 강화학습은 결과에 대한 피드백을 제공하고 정확한 입력값과 출력값은 제공하지 않는 학습방법이다[12]. 본 연구에서는 지도학습 방법인 회귀모델을 이용하여 조회수를 예측하고, 관련성 랭킹 예측, 구독자수 등을 예측한다.

2-3 딥러닝

다계층적 구조의 신경망을 가진 모델로서 딥러닝(deep learning)은 데이터를 표현하거나 입력과 출력 사이의 관계를 충분히 설명하기 위한 더 복잡하고 깊은 신경망을 가진 머신러닝에 적합한 특징을 추출하는 알고리즘이다. 복잡도에 있어 기존의 천층망(shallow network)과 딥러닝의 심층망이 유사한 형태를 보이지만 심층망에서는 입력과 출력 사이 경로를 더 다양한 방법으로 모델링할 수 있다는 특성이 있다. 또한 음성인식과 영상인식에서 천층망이 인간 두뇌의 성능에 미치지 못한다는 점에 비해 심층망은 생물학적 신경망의 원리를 이용한 계층적 모델을 적용한 것이다. 다계층적 심층망은 전문가의 지식에 의존하지 않고 자동으로 데이터의 특징을 추출가능하게 되었고, 분류기와 특징 추출을 통합, 학습하는 성능이 진일보하였다[12],[14].

딥러닝 기술이 적용된 신경망 구조는 자가인코더, 제한볼츠만기계, 합성곱신경망(convolutional neural network: CNN), 회귀신경회로망(recurrent neural network: RNN) 등이 있다. 이 중 영상인식 분야는 CNN[14], 음성음식 등 시계열 데이터처리에 RNN이 주로 사용된다[12],[14].

CNN은 뇌신경 과학적 발견들에 기초하여 “지역적 감각수용장을 전체 이미지에 합성곱(convolution)하여 신경망의 연결강도를 공유하는 방법으로 변수의 개수를 대폭 축소‘한 모델이다. RNN은 시계열 데이터에 적합한 구조로 일반 신경망에서 각 계층을 상위계층으로 연결하고 자기 계층에서도 연결함으로써 연결이 메모리와 같은 역할을 하게 되는 것이다. 이를 통해 데이터의 시간적인 변화에 대해 모델링이 가능해진다. 최근 장단기기억모델(LSTM)등을 사용하면서 돌아오는 연결로 학습이 더 어려운 학습문제를 해소하며 널리 사용되고 있다[14]. 본 연구에서는 예측 알고리즘 구현을 위해 CNN 딥러닝 알고리즘으로 예측정확도를 확인한다.

2-4 랜덤 포레스트와 변수중요도

랜덤 포레스트는 초기 Yali Amit & Donald Geman(1996)[15]과 Tin Kam Ho(1998)[16]의 연구 영향과 Breiman(2001)이 제시한 랜덤 노드 최적화(randomized node optimization)와 배깅(bagging)을 동시에 사용한 CART(classification and regression tree)를 이용하고 상관관계가 없는 트리들로 구성한 포레스트 방식이다[17].

랜덤 포레스트 구조는 1개 트리가 노드(node)와 에지(edge)의 집합으로 구성되며 계층구조를 이룬다. 랜덤 포레스트는 다수의 의사결정트리(Decision Tree)들을 학습하는 방법으로 학습데이터를 독립변수로 한 의사결정트리 구조를 자동 생성하는 과정을 통해 분류와 예측 문제를 해결하는 모델이다. 여기에서 의사결정트리(decision tree)는 클래스 또는 레이블로 된 것을 분류(classification)하는 것을 말하며, 회귀트리(regression tree)는 종단노드(terminal 또는 leaf node) 데이터의 평균값을 사용하여 결과값을 추정하는 것을 말한다.

랜덤 포레스트는 변수중요도(variable importance)를 사용하여 결과를 해석하며[18], 접근 방식은 Feature importance와 Permutation feature importance이다.

Feature importance는 주어진 모델에서 feature가 예측결과에 영향을 미치는 정도를 측정한다. 다양한 feature에 기반하여 예측 모델이 구축된다는 점에서 각 feature는 고유성을 지닌다. 각 feature는 독립적으로 평가되며 feature 간의 상호 작용은 무시된다[19]. Permutation importance(순열중요도)는 예측 변수인 feature 중 하나를 랜덤하게 순열한 후 그 성능을 측정하고 예측 오차 증가량에 따라 feature별 중요도를 측정하는 방법이다[20]. 어떤 feature가 중요한 역할을 하는 경우, 그 feature를 제거한 이후에 예측 오차가 증가한다고 해석할 수 있다[19]. 본 연구에서는 상기 2가지 방법을 사용하여 입력변수와 출력변수에 대한 변수중요도를 도출한다.

Ⅲ. 연구방법

본 연구는 유튜브 사이트에 등록된 영상들의 데이터를 검색어를 통해 추출하여 머신러닝을 통해 관련성, 조회수, 채널 구독자수와 상관관계가 높은 변인들을 도출하고, 머신러닝과 딥러닝을 통해 예측 알고리즘까지 구현해 보았다. 연구는 아래의 순서에 따라 진행되었다.

3-1 연구 문제의 설정

크리에이터들은 영상 조회수와 채널 구독자수를 늘리기 위해 노력한다. 조회수와 채널 구독자수는 어떤 변인들과 상관관계가 높은지 확인하고자 한다. 그리고 조회수와 채널 구독자수를 늘리기 위해서는 검색 시 영상추천이 상위에 랭크되어야 하는데 기본 필터링 기준인 관련성 기준은 어떤 변인들과 상관관계가 높은지 확인하려고 한다.

연구문제 1: 유튜브 검색을 통해 추천되는 영상들은 관련성 정렬기준으로 어떤 변수들과 상관관계가 높은가?

연구문제 2: 유튜브 영상의 조회수는 어떤 변수들과 상관관계가 높은가?

연구문제 3: 유튜브 채널 구독자수는 어떤 변수들과 상관관계가 높은가?

3-2 데이터 추출과 전처리

유튜브 영상에 관련된 다양한 정보들 중 본 연구에서는 유튜브 플랫폼에서 제공하는 영상의 정보들을 API 연동을 통해서 데이터 크롤링 기법으로 추출하였다. 소프트웨어는 Python 3.9 버전을 사용하였다. 검색어, 정렬기준, 지역, 정보 추출량 정보를 입력하면 데이터를 추출할 수 있다. 유튜브는 분류번호 기준으로 영상의 종류를 총 32가지로 분류하는데 Music, Pet & Animals, People & Blogs, Entertainment, News & Politics, Education의 6가지 카테고리와 관련된 검색어를 이용하였다. 지역은 한국으로 하였고, 정보추출은 1회 최대 제공기준인 50개를 여러 검색어를 통해 반복 추출하였다. 기간은 2021년 12월 1개월간 검색어 기준 16,194개, 조회수 기준 8,200개의 데이터를 추출하였다.

표 1. 변인의 추출과 가공
Table 1. Extracted and processed variables

Variables	Method	Select
Relevance ranking	extracted	O
Number of likes	extracted	O
Number of dislikes	extracted	O
Number of views	extracted	O
Preferred registration	extracted	X
Number of comments	extracted	O
Category	extracted	X
Registration time	extracted	X
Exposure periods	processed	O
Title of videos	extracted	X
Number of words_Title	processed	O
Frequency of searching word_Title	processed	O
Channel title	extracted	X
Number of subscribers	extracted	O
Channel ID	extracted	X
Video running time	extracted	O
Video description	extracted	X
Number of words_Description	processed	O
Frequency of searching word_Description	processed	O
Video tag	extracted	X
Number of words_Tag	processed	O
Frequency of searching word_Tag	processed	O

유튜브에서 추출한 변인들은 총 15개이며, 해당 데이터로 가공하여 7개의 변인들을 추가로 가공하였다. 수집한 데이터 중 추가 가공 후 불필요한 8개는 제외하고 총 14개의 변인들을 선정하였고 내역은 Table 1과 같다.

채널 ID 기준으로 중복데이터 제거, 결측 데이터는 0으로 채워 총 12,664개의 영상데이터 정보를 확보하였다. 수집한 영상데이터는 조회수는 0~10억뷰, 채널구독자수도 0~1억명으로 넓은 구간에 걸쳐 우하향 롱테일을 가지는 특징을 나타내었다. 머신러닝과 딥러닝의 예측의 정확도를 높이기 위해서는 상위 30% 이상 구간의 롱테일 데이터를 삭제하였다.

3-3 데이터 분석

우선 머신러닝을 활용한 상관관계 분석을 위해 Decision Tree Regression 방법을 이용하여 머신러닝을 진행하였고, Feature Importance와 Permutation Importance의 방법을 통해 입력변수들의 출력변수에 대한 변수중요도를 도출하였다.

아울러 머신러닝과 딥러닝을 활용한 예측 알고리즘 구현을 위해 DTR(Decision Tree Regression) 머신러닝 알고리즘과 CNN(Convolutional Neural Network) 딥러닝 알고리즘을 이용하여 조회수와 채널구독자수 예측 프로그램을 구현하여 예측 정확도를 확인하였다.

Ⅳ. 연구결과 및 분석

4-1 상관관계 분석

머신러닝의 Feature Importance와 Permutation Importance를 이용하여 유튜브 검색의 관련성 필터기준, 영상의 조회수, 채널의 구독자수와 상관관계가 높은 변인들을 확인하였다.

첫째, 유튜브 관련성 기준 검색 추천 영상들의 순위(상위 50개 영상의 정보만 추출 가능)와 추출변인들과의 상관관계를 확인하였다. 변인들 중 조회수, 노출기간, 영상길이만이 상관관계가 있는 것으로 나타났으며 조회수가 가장 관련이 높게 나타났다(Fig. 1).

유튜브 검색시 기본 필터기준인 관련성 기준으로 영상을 표출할 때의 알고리즘은 공개되지 않았다. 실제로 추출이 가능한 다양한 데이터 중 머신러닝 기법으로 확인하여도 3가지 정도의 변인들만이 결과에 영향이 있었는데 조회수가 높거나 최근에 올라온 영상이 더 상위에 제시될 가능성이 높다. 영상의 길이를 길게 만드는 것도 검색 표출순위를 높이는데 도움이 된다.

둘째, 조회수에 관련성이 높은 변인들을 확인하였다(Fig. 2). 좋아요수, 싫어요수가 가장 상관관계가 높았고, 댓글, 노출기간, 제목에 검색어 빈도수, 영상길이가 그 다음으로 상관관계가 높은 것으로 나타났다.

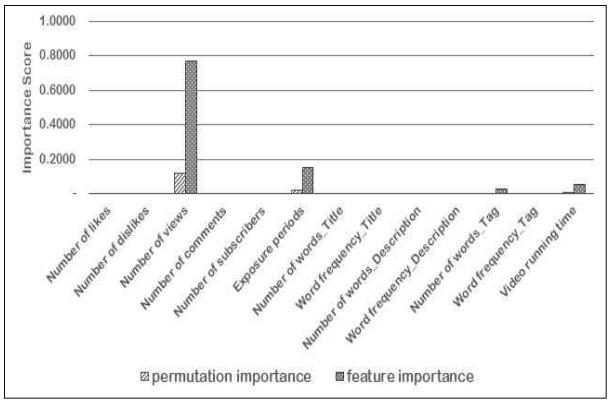

그림 1. 검색 추천 영상 변인들의 순위
Fig. 1. Importance scores of variables to the relevance

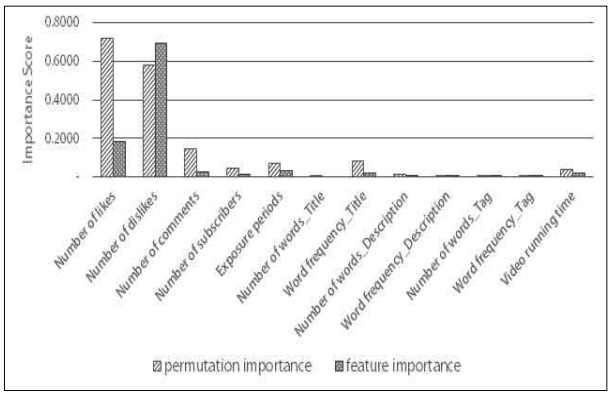

그림 2. 조회수에 관련성이 높은 변인들의 순위
Fig. 2. Importance scores of variables to the view

구독자 수는 상관관계가 낮은 특징을 나타내었다.

조회수는 시청자의 영상에 대한 좋아요, 싫어요, 댓글과 같은 적극적인 참여가 조회수를 높이는데 효과가 있다. 영상 중간 멘트나 자막 등을 통해 시청자의 참여를 유도하는 활동이 조회수를 높이는데 효과적이라고 판단된다.

조금 더 구체적으로 영상 카테고리에 따라 조회수에 관련성이 높은 변인들의 특징에 차이가 있는지 살펴보았다(Fig. 3). 조회수에 관련성이 높은 좋아요, 싫어요, 댓글의 3가지 변인들에 대한 관련성 점수를 비교하였다. 음악과 엔터테인먼트의 경우 좋아요에 관련성이 높게 나왔는데, 음악이나 엔터테인먼트 영상을 볼 경우 마음에 들면 좋아요로 반응을 하고 싫어요를 주도적으로 누르지 않을 것으로 판단되고 이는 조회수에 대한 상관성 결과와도 연결되는 것으로 판단된다. 인물&블로그, 뉴스&정치의 경우 싫어요에 민감하게 반응하는 것으로 나타났으며 뉴스&인물의 경우 댓글에도 민감한 것으로 보인다. 자극적인 이슈들이 더 관심을 끄는 일반적인 현상을 그대로 보여주는 결과라고 할 수 있다. 교육의 경우 좋아요, 싫어요, 댓글 모두가 균형있게 높은 점수를 보이는데, 교육 영상을 조회하는데 다양한 정보들을 비중있게 보고 판단한다고 유추할 수 있다.

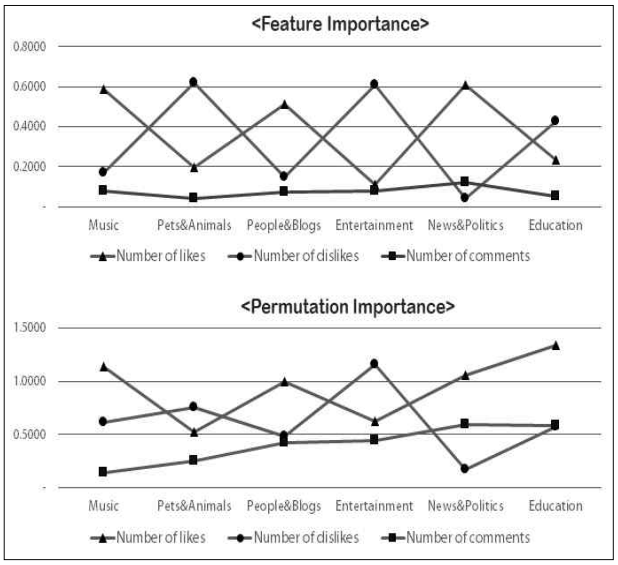

그림 3. 조회수에 관련성이 높은 카테고리 순위
Fig. 3. Importance score by Category

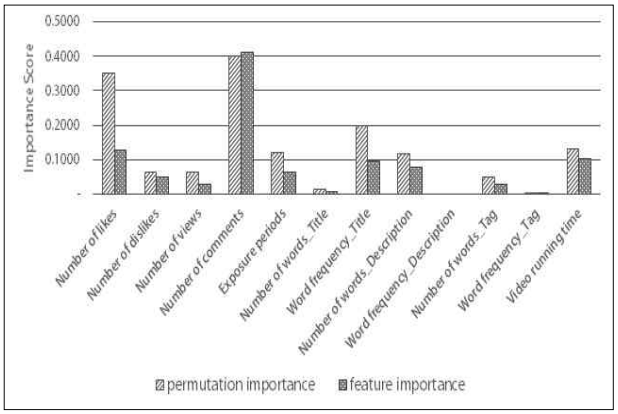

그림 4. 구독자수에 관련성이 높은 변인들의 순위
Fig. 4. Importance scores of variables to the Subscriber

셋째, 구독자수에 관련성이 높은 변인들을 확인하였다(Fig. 4). 좋아요수, 댓글수가 상관관계가 높았고, 노출기간, 제목에 검색어 빈도수, 영상설명 단어수, 영상길이가 그 다음으로 상관관계가 높게 나타났다. 싫어요수나 조회수는 상관관계가 낮은 특성을 보였다.

구독자를 늘리기 위해서는 시청자의 좋아요나 댓글과 같은 긍정적인 피드백이 필요하다. 좋은 영상을 본 후 구독을 하게 되므로 영상 자체의 내용과 품질이 좋아야 한다는 것을 알 수 있다.

4-2 머신러닝과 딥러닝을 이용한 예측 프로그램 구현

DTR 머신러닝 알고리즘을 이용한 조회수 및 구독자수 예측 프로그램(Fig. 5)과 CNN 딥러닝 알고리즘을 이용한 조회수 및 구독자수 예측 프로그램(Fig. 6)을 구현하여 예측 정확도를 확인하였다.

```
#8:2로 train data와 test data 분류
from sklearn.model_selection import train_test_split
x_train1, x_test1, y_train1, y_test1=train_test_split(x1, y1, test_size=0.2, random_state=2021)

#DecisionTreeRegressor로 머신러닝
from sklearn.tree import DecisionTreeRegressor
#최적의 max_depth 찾기
trees=[]
for depth in range(1, 20):
    tree=DecisionTreeRegressor(max_depth=depth, min_samples_leaf=10, random_state=2021)
    tree.fit(x_train1, y_train1)
    trees.append(tree)
eccs=[]
for depth in range(1, 20):
    idx=depth-1
    tree=trees[idx]
    ecc=tree.score(x_test1, y_test1)
    eccs.append(ecc)
#최적의 max_depth를 그래프로 확인
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.plot(range(1, 20), eccs)
plt.grid()
plt.show()

#동일한 방식으로 최적의 min_sample_leaf 찾기

# (중략/생략)

#최적 모델 학습 (둘 다 7이 최적의 값)
model=DecisionTreeRegressor(max_depth=7, min_samples_leaf=7, random_state=2021)
model.fit(x_train1, y_train1)

#예측과 실제 결과 그래프로 비교
y1=model.predict(x_train1)
plt.scatter(y_train1, y1)
plt.show()

#train과 test MAE 값 확인
MAE(y_train1, y1)
MAE(y_test1, y)
```

그림 5. DTR 조회수 및 구독자수 예측 프로그램 소스코드
Fig. 5. Views & Subscriber prediction source code of DTR

```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.layers import Input, Dense, BatchNormalization
from tensorflow.keras.models import Sequential, Model
from tensorflow.keras.callbacks import EarlyStopping

keras.backend.clear_session()

#딥러닝모델 구현
model=Sequential()
model.add(Dense(24, activation='relu', input_shape=(12,))) #입력변수 12개
model.add(BatchNormalization())
model.add(Dense(48, activation='relu'))
model.add(BatchNormalization())
model.add(Dense(48, activation='relu'))
model.add(BatchNormalization())
model.add(Dense(24, activation='relu'))
model.add(BatchNormalization())
model.add(Dense(1, activation='relu'))
model.compile(optimizer='adam', loss='mse')
es=EarlyStopping(monitor='val_loss', min_delta=0,
                 patience=100, verbose=1, restore_best_weights=True)
history=model.fit(x_train1, y_train1, epochs=2000, batch_size=32,
                  verbose=1, validation_data=(x_test1, y_test1), callbacks=[es])

#결과값 예측
y11=model.predict(x_test1)
y12=model.predict(x_train1)

#train과 test의 MAE 값 확인
MAE(y_train1, y12)
MAE(y_test1, y11)
```

그림 6. CNN 조회수 및 구독자수 예측 프로그램 소스코드
Fig. 6. Views & Subscriber prediction source code of CNN

먼저 영상의 조회수를 예측하는 DTR과 CNN 예측모델은 양호한 예측 결과를 나타내었다(Fig. 7). 예측정확도를 측정하는 MAE(mean absolute error) 및 MSE(mean square error) 값은 CNN의 경우 표준화된 변수값을 사용하여 두 예측모델의 정확도를 1:1로 비교하지는 못하였으나 그래프에서도 볼 수 있는 것처럼 CNN이 조금 더 나은 결과를 나타내고 있다. 다음으로 채널의 구독자수를 예측하는 DTR과 CNN 예측모델도 좋은 예측 결과를 나타내지 못하였다(Fig. 8). 채널과 관련하여서는 추출된 변수 이외에도 다른 변인들에 대한 추가적인 확인이 필요할 것으로 판단된다.

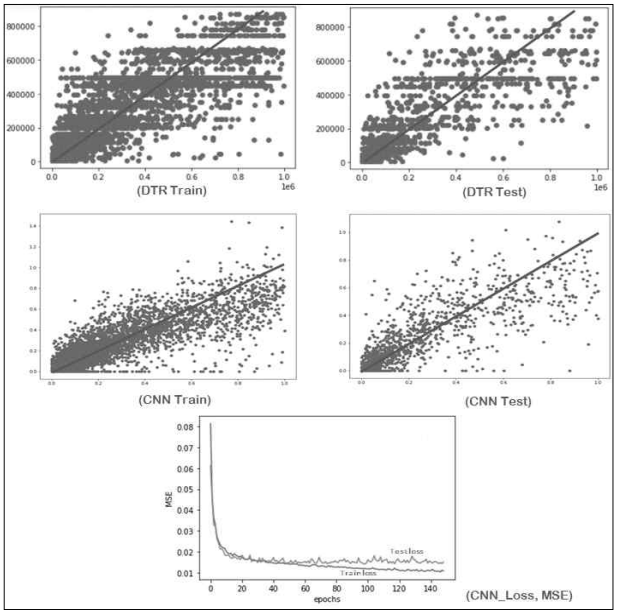

그림 7. DTR과 CNN에 의한 조회수 예측
Fig. 7. Views prediction result of DTR and CNN

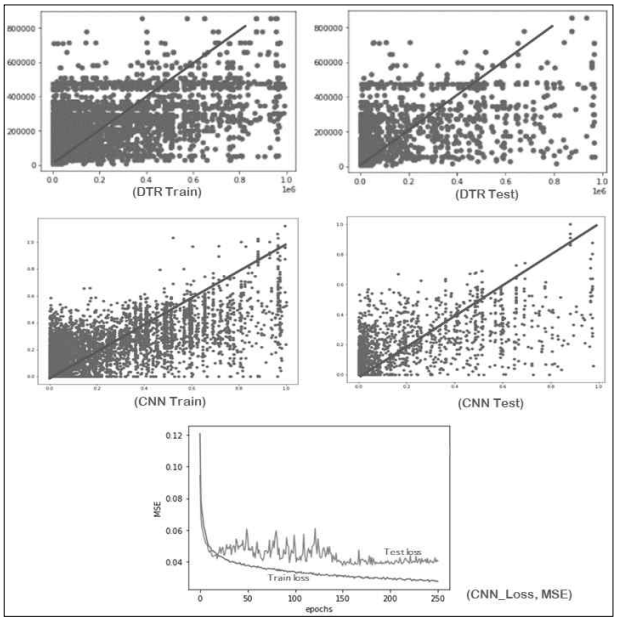

그림 8. DTR과 CNN에 의한 구독자수 예측
Fig. 8. Subscribers prediction result of DTR and CNN

조회수와 채널구독자수는 0~10,000,000 정도의 넓은 구간에 걸쳐 분포되어 있어 예측 정확도는 많이 떨어졌다. 정확도가 높은 결과를 얻기 위해서는 조회수와 채널 구독자수가 적은 구간(초기 단계)과 많은 구간(성숙 단계) 별도로 구성하여 예측을 별도로 하거나, 구간을 설정하여 구간 예측을 통해 예측 개선이 되는지 확인해 볼 필요성도 있다. 딥러닝에서 권장하고 있는 10만개 이상의 많은 데이터를 이용하여 예측을 진행할 필요성도 있다.

V. 결 론

본고는 유튜브에 조회수와 구독자수와 연관이 높은 변인들, 영상 검색 시 상위노출과 관련이 높은 변인들이 무엇인지 확인하고자 하였다.

연구결과 상관관계 분석에서 변수의 중요도가 매우 중요하다는 결론을 도출했다. 즉, 변수의 중요도는 전체적인 변수와 통제 가능한 변수로 구분할 수 있다. 첫째, 좋아요/싫어요는 전체적으로 좋아요 수가 조회수에 가장 영향을 미치고 있으며, 다른 변수들에 비해 조회수에 영향도가 더 높은 변수임을 확인했다. 둘째, 구독자 수가 조회수에 미치는 영향은 거의 없다. 셋째, 영상 태그 단어 수/영상 길이에는 상세한 설명을 제시(단어 수가 많음) 할수록 조회수가 높았으며, 영상의 길이도 조회수에 높은 변수임을 재확인 했다. 머신 러닝과 딥 러닝을 이용한 예측 프로그램의 분석에서는 데이터의 분포가 넓어 조회수 예측의 정확도는 많이 떨어졌으나 학습 데이터, 예측 데이터는 머신 러닝과 딥 러닝 중 어느 쪽을 채택했을 때 정확도가 더 높은 성능을 보여줄 수 있는지 확인할 수 있었다.

아울러 본연구의 함의는 다음과 같다. 영향도에 대한 변수 확인은 딥러닝에서는 직접적으로 확인이 어렵고, 머신 러닝의 Feature Importance 및 Permutation Importance를 통해 주요 변수를 확인할 수 있었다. 조회수에 영향을 미치는 변수로는 좋아요 수, 싫어요 수, 댓글 수, 노출기간 등으로 관계가 높을 것으로 예측한 구독자 수는 조회수에 영향도가 낮게 나타났다. 그리고 예측 정확도는 딥 러닝이 일반적으로 알려진 것처럼 정확도가 조금 더 높게 나타났고, 학습 데이터는 머신 러닝이 더 정확했다.

본 연구는 상기의 연구 성과에도 불구하고 다음과 같은 한계점을 지니고 있다. 유튜브 검색 시 검색결과 순위의 또 다른 중요 필터 기준인 관련성에 대한 추가적인 연구가 필요하며, 구독자 수가 조회수에 영향이 낮은 이유에 대한 추가적인 분석도 필요하다. 또한 유튜브의 썸네일이 중요한 역할을 하고 있는데 이에 대한 세부적인 썸네일의 추가적인 분석이 필요할 것이라고 판단된다.

본 연구를 통하여 유튜브의 조회수에 어떠한 변수들이 상관관계가 높은지, 유튜브 영상의 카테고리에 따라 상관성이 높은 변수들에 차이가 있는지를 확인할 수 있다. 이를 통해 유튜버들은 영상의 조회수를 높이고 구독자를 늘리기 위해 어떠한 부분에 더 주의를 기울여야 하는지를 파악할 수 있을 것이며, 콘텐츠 이용자는 어떠한 부분에 더 주의를 기울이는지 파악할 수 있어, 더 나은 고객서비스 환경을 만들어 갈 수 있을 것으로 기대된다.

감사의 글

이 논문은 2021년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임
(NRF-2021-R1I1A3054903).

참고문헌

[1] Youtube Support Page. http://support.google.com/youtube/.

[2] H. S. Kim, "Analysis of Popular YouTube Video Content using Data Mining," *Journal of Digital Contents Society*, Vol. 21, No. 4, pp. 673-681, April 2020. https://doi.org/10.9728/dcs.2020.21.4.673

[3] S. J. Lee and S. B. Lee, "Diffusion Strategies for K-Beauty Hallyu Contents on YouTube," *GRI Review*, Vol. 20, No. 3, pp. 231-259, August 2018.

[4] Y. S. Lim, "Exploring the Direction for Violence Prevention Campaign Using YouTube : Focusing YouTube Video Network Analysis," *Advertising Research*, Vol. 124, pp. 65-100, March 2020. https://doi.org/10.16914/ar.2020.124.65

[5] J. S. Lee and A. Nerghes, "Refugee or Migrant Crisis? Labels, Perceived Agency, and Sentiment Polarity in Online Discussions," *Social Media+ Society*, Vol. 4, No. 3, July 2018. https://doi.org/10.1177/2056305118785638

[6] J. B. Schmitt, D. Rieger, O. Rutkowski and J. Ernst, "Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube: Recommendation Algorithms," *Journal of Communication*, Vol. 68, No. 4, pp. 780-808, August 2018. https://doi.org/10.1093/joc/jqy029

[7] S. Y. Yoo and O. R. Jeong, "The YouTube Video Recommendation Algorithm using Users, Social Category," *Journal of KIISE,* Vol. 42, No. 5, pp. 664-670, May 2015. https://doi.org/10.5626/JOK.2015.42.5.664

[8] K. M. Kang, J. W. Eom, H. J. Jung, S. W. Park, J. S. Lee, H. M. Kang and T. W. Kang, "Classification System for Recommending YouTube Popular Videos," *Proceedings of KIIT Conference,* pp. 595-597, October 2020.

[9] S. H. Lee, *Return of Barbarism, YouTube Reality and Prospect*, Yerinwon, September 2020.

[10] S. J. Bae, "Trend Analysis of Movie Content Curation and Metadata Standards Research : Focus on the Art Management Perspective," *Journal of the Korea Convergence Society*, Vol. 11, No. 6, pp. 163-171, June 2020. https://doi.org/10.15207/JKCS.2020.11.6.163

[11] S. J. Bae and S. H. Lee, "A Study on the Curation Factors through Reverse Engineering Design of YouTube

Algorithm - Focusing on Gender Keyword Search," *Journal of the Korea Convergence Society*, Vol. 13, No. 3, pp. 133-146, March 2022. https://doi.org/10.15207/JKCS.2022.13.03.133

[12] S. E. Mun, S. B. Jang, J. H. Lee and J. S. Lee, "Machine Learning and Deep Learning Technology Trends," *Information and Communications Magazine*, Vol. 33, No. 10, pp. 49-56, October 2016.

[13] J. H. Ku, "A Study on the Machine Learning Model for Product Faulty Prediction in Internet of Things Environment," *Journal of Convergence for Information Technology*, Vol. 7, No. 1, pp. 55-60, February 2017. https://doi.org/10.22156/CS4SMB.2017.7.1.055

[14] H. Y. Choe and Y. H. Min, "Introduction to Deep Learning and Major Issues," *Korea Information Processing Society Review,* Vol. 22, No. 1, pp. 7-21, January 2015.

[15] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," *Neural Computation*, Vol. 9, No. 7, pp. 1545-1588, October 1997. https://doi.org/10.1162/neco.1997.9.7.1545

[16] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832-844, August 1998. https://doi.org/10.1109/34.709601

[17] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, October 2001. https://doi.org/10.1023/A:1010933404324

[18] S. B. Shin and H. J. Cho, "Correlated variable importance for random forests," *The Korean Journal of Applied Statistics*, Vol. 34, No. 2, pp. 177-190, April 2021. https://doi.org/10.5351/KJAS.2021.34.2.177

[19] S. J Oh, "Predictive Case-based Feature Importance and Interaction," *Information Sciences*, Vol. 593, pp. 155-176, February 2022. https://doi.org/10.1016/j.ins.2022.02.003

[20] D. H. Yoon, M. T. Yoo, J. J. Park, K. H. Kim, M. J. Lee and I. W. Lee, "Analysis of Accuracy in Predicting Stability of Piers Evaluated by Impact Load Test," *Journal of the Korean Society for Railway*, Vol. 24, No. 7, pp. 581-589, July 2021. https://doi.org/10.7782/JKSR.2021.24.7.581

김철년(Chul-Nyuon Kim)

2011년 : 연세대학교 정보대학원 (이학석사)
2017년 : 경성대학교 디자인전문대학원 (미디어학박사)

2002년～현 재: KT
※관심분야 : 인공지능, 커뮤니케이션, 알고리즘 등

배승주(Seung-Ju Bae)

2007년 : 부산대학교 예술문화영상매체 (예술학석사)
2017년 : 부산대학교 예술문화영상매체 (예술학박사)

2019년～현 재: 경성대학교 미디어콘텐츠학과 외래교수
※관심분야 : 예술경영, 문화콘텐츠, 미디어, 큐레이션 등

하윤수(Youn-Soo Ha)

2006년 : 한양대학교 경영학과 (경영학석사)
2021년 : 경성대학교 디자인전문대학원 (미디어학박사)

2020년～현 재: 허브스코프
※관심분야 : 디지털마케팅, O2O, OMO, 인공지능, 알고리즘 등

이상호(Sang-Ho Lee)

2003년 : Aalto University (경영학석사)
2008년 : 서울과학종합대학원 (경영학박사)

2010년～현 재: 경성대학교 미디어콘텐츠학과 교수
※관심분야 : 미디어, 마케팅, 콘텐츠, 알고리즘 등