삼중항 손실 기반의 조인트 임베딩을 이용한 영상 콘텐츠 유사도 측정 방법

홍택은, 김판구*
조선대학교 컴퓨터공학과
e-mail : goodfax2000@naver.com, pkkim@chosun.ac.kr*

A Method of Video Content Similarity Measurement using Triplet Loss-based Joint Embedding

Taekeun Hong, Pankoo Kim*
Department of Computer Engineering, Chosun University

요 약

스마트 기기의 보유율이 꾸준히 증가와 코로나 여파로 인해 동영상제공서비스(OTT) 이용율이 크게 증가했다. OTT 이용 만족도와 지속 사용 의도를 결정하는 요소 중 추천 시스템의 만족도가 중요하다. 그러나 대부분의 추천 시스템은 사용자에게 메타 데이터를 강요하고 화제성 높은 콘텐츠를 추천한다. 콘텐츠 자체를 분석하여 서비스를 제공하는 것이 필요하다. 텍스트와 이미지를 함께 분석하는 방법으로 조인트 임베딩이 주로 쓰이지만 콘텐츠를 도메인으로 하는 연구는 미비하다. 따라서 본 논문에서는 텍스트와 이미지로 구성된 콘텐츠를 분석하기 위해 삼중항 손실 기반의 조인트 임베딩을 이용하여 영상 콘텐츠의 유사도를 측정한다. 데이터로 영화 줄거리와 스틸컷, 포스터를 사용한다. 텍스트 임베딩에 KoBERT를 사용하고 이미지 특징 추출에 EfficientNet을 이용했으며, 이를 결합한 뒤 삼중항 손실을 이용항 영화 콘텐츠의 거리를 학습했다. 그 결과 영화의 조인트 임베딩을 성공적으로 수행했으며, 실제 영화 정보를 통해 유사함을 확인하고 비교하여 타당함을 확인했다. 본 논문에서 제안한 방법을 통해 사용자가 시청한 콘텐츠와 유사한 콘텐츠를 추천 할 때 사용 가능할 것으로 보인다.

1. 서 론

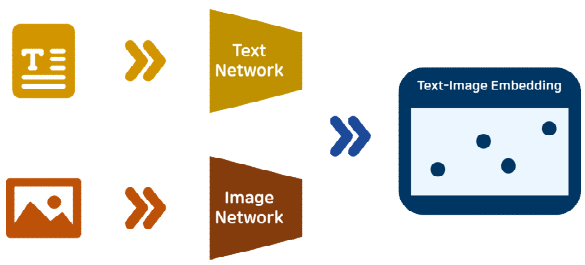
스마트폰과 태블릿PC와 같은 스마트 기기의 보유율이 꾸준히 증가하고 있으며, 코로나의 여파로 인해 스마트 기기를 이용한 미디어 콘텐츠 이용률과 이용 시간도 크게 증가하고 있다[1]. 특히 동영상제공서비스(OTT)를 이용하는 비율이 크게 증가했으며, 주로 YouTube, NETFLIX, 왓챠, Wavve, NAVER TV 등을 이용하는 것으로 나타났다[2]. OTT 사용자의 증가에 따라 이용 만족도와 지속 사용 의도에 미치는 요소를 분석하는 것이 중요해졌다. 중요한 요소로 콘텐츠 다양성, 요금제 적절성, 추천 시스템, N스크린 서비스, 몰아보기 기능 등이 있으나 추천 시스템이 사용자 만족도 및 지속 사용 의도에 가장 중요한 요소로 작용한다[3]. 그러나 대부분의 추천 시스템은 사용자의 메타 데이터를 강요하여 추천 서비스를 제공하거나 화제성이 높은 콘텐츠를 추천한다는 문제점이 존재한다. 그렇기 때문에 콘텐츠를 분석하여 사용자가 주로 즐기는 콘텐츠와 유사한 콘텐츠를 추천하는 것은 중요하다. 텍스트와 이미지 정보를 함께 분석하기 위한 방법으로 조인트 임베딩 방법이 활용된다. 일반적인 조인트 임베딩은 텍스트와 이미지를 결합하기 위한 임베딩 방법으로 임베딩 된 도메인들 사이의 관계를 잘 표현하는 것을 목표로 하며, 이미지-텍스트 검색, 이미지 캡셔닝, Visual Question Answering(VQA)에 활용한다[4]. 하지만 텍스트와 이미지로 구성된 콘텐츠를 분석하는 연구는 미비한 실정이다. 따라서 본 논문에서는 삼중항 손실 기반의 조인트 임베딩을 이용한 영상 콘텐츠 유사도 측정 방법을 제안한다. 콘텐츠의 텍스트 정보와 이미지 정보를 결합하는 임베딩 방법을 통해 영상 콘텐츠의 유사도를 측정하고 사용자에게 추천한다.

본 논문의 구성은 다음과 같다. 2장에서 일반적인 조인트 임베딩 방법과 삼중항 손실에 대해 기술하며, 3장에서는 삼중항 손실 기반의 조인트 임베딩을 이용한 영상 콘텐츠 유사도 측정 방법에 대해 기술한다. 마지막으로 4장에서 결론 및 향후연구로 마무리한다.

2. 관련 연구

2.1 조인트 임베딩

조인트 임베딩은 텍스트와 이미지를 각 딥러닝 네트워크를 통해 벡터로 변환하고 결합하는 방법으로 이미지-텍스트 검색, 이미지 캡셔닝, VQA에 주로 활용된다.



(그림 1) 일반적인 조인트 임베딩의 구성

조인트 임베딩은 그림 1과 같이 텍스트와 이미지의 특징을 추출하기 위한 각 딥러닝 네트워크로 구성된다[4, 5, 6, 7]. 일반적으로 텍스트 네트워크는 텍스트 임베딩을 위한 방법으로 구성되며, Word2Vec[8], Recurrent Neural Network(RNN)[9], Long Short-Term Memory(LSTM)[10], BERT[11]로 구성될 수 있다. 이미지 네트워크는 이미지 특징을 추출하기 위한 방법으로 ResNet[12], EfficientNet[13], Inception[14] 등의 특징 추출기로 구성될 수 있다. 마지막으로 텍스트와 이미지를 임베딩하는 과정은 삼중항 손실이나 분류기로 구성될 수 있다.

2.2 삼중항 손실

삼중항 손실은 학습에 3개의 데이터를 이용하며, 기준이 되는 데이터를 Anchor, 나머지 2개의 데이터를 Negative와 Positive라고 정의한다. Anchor와 Positive, Negative의 유사도를 측정하여 Positive는 더욱 가깝게 학습하고 Negative는 더욱 멀게 학습하는 것이 목표이다. 삼중항 손실의 수식은 아래와 같다.

$$Loss = \max(0, m + d(f(x_a), f(x_p)) - d(f(x_a), f(x_n)))$$

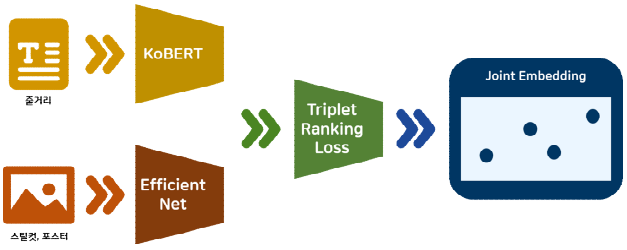
여기서, x_a는 anchor이고 x_p는 positive, x_n은 negative, m은 마진이다. 따라서 anchor와 negative의 거리 값이 anchor와 positive의 거리 값보다 크면 정답이기 때문에 loss를 0으로 업데이트한다. 반대로 anchor와 positive의 거리 값이 anchor와 negative의 거리 값보다 크면 두 거리 값을 뺀 값으로 loss를 업데이트 해가면서 정답이 될 수 있게 학습한다[15].

3. 삼중항 손실 기반의 조인트 임베딩을 이용한 영상 콘텐츠 유사도 측정 방법

본 장에서는 본 논문에서 제안한 방법과 실험 결과를 설명한다. 실험에 사용한 데이터는 영화를 도메인으로 하며 네이버 영화에서 수집했다. 텍스트는 줄거리로 활용하고 이미지는 스틸컷과 포스터를 사용했다. Anchor에 해당하는 영화는 2,000개이고 Positive와 Negative에 해당하는 영화는 각각 3,489개이다.

조인트 임베딩에서 텍스트 임베딩을 위한 네트워크에 KoBERT[11]를 사용했고 이미지 특징 추출을 위한 네트워크에 EfficientNet[13]을 사용했다. 추출한 텍스트 임베딩과 이미지 특징은 벡터 합 연산을 통해 결합하고 학습을 위해 삼중항 손실을 이용한다.

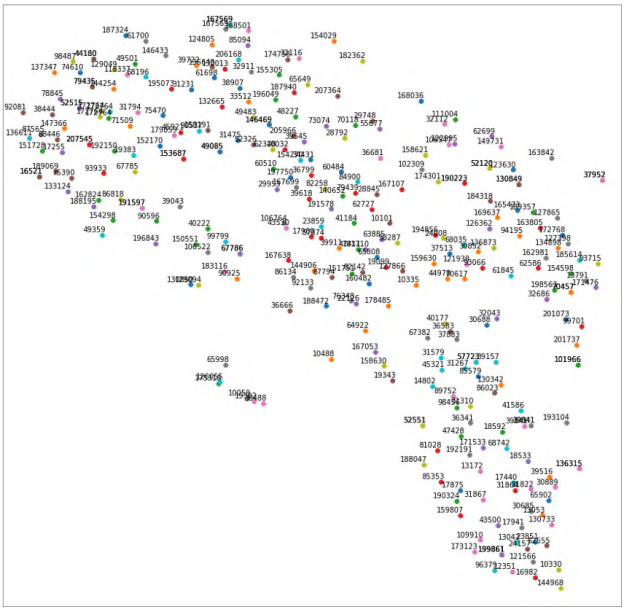
그림 2는 본 논문에서 제안한 삼중항 손실 기반 조인트 임베딩의 구성이다. KoBERT와 EfficientNet은 사전학습 된 모델을 이용했으며, lookup table과 같은 형태로 입력 데이터를 추출했다. KoBERT의 출력 임베딩 크기는 768이다. EfficientNet을 B7을 이용했고 출력 특징 크기는 1,000이지만 KoBERT와 출력 크기를 맞추기 위해 768로 조절했다.



(그림 2) 삼중항 손실 기반 조인트 임베딩의 구성

Tiplet Ranking Loss 단계에서 1차원 Convolution과 1차원 MaxPooling, Dropout으로 구성된 모듈을 6개 이용했으며, 그 뒤에 Fully Connected(FC) 레이어와 L2정규화를 수행한다. 각 Convolution의 활성화 함수는 ReLU이고 MaxPooling의 필터 크기는 2, Dropout rate는 0.3이다. 학습에 사용한 optimizer는 Adam이고 Learning Rate는 0.001, batch size는 32이다. 학습 데이터 셋과 테스트 데이터 셋은 8:2로 구성하여 실험했다. 그 결과 도출되는 조인트 임베딩의 출력 크기는 768이다. 그 후 T-Stochastic Neighbor Embedding(TSNE)을 이용하여 임베딩 크기를 2로 축소하여 시각화했다.

테스트 데이터로 영화 8,959개를 사용했으며, 그림 3은 테스트 데이터로 조인트 임베딩한 것 중 영화 300개에 대한 결과이다.



(그림 3) 조인트 임베딩의 시각화 결과

시각화 결과 중에서 거리가 가까운 것으로 보이는 영화를 실제로 유사한지 직접 확인하여 사람의 관점에서도 유사한지 살펴보았다. 영화 32112와 11004, 158630과 19343, 187324와 61700이 각각 유사한 것으로 나타났다. 32112와 11004는 애니메이션으로 주인공이 모험을 떠나는 내용을 가진 영화다. 158630과 19343은 해외 영화이고 드라마, 로

맨스 코미디로 일상생활에서 발생하는 남녀의 이야기를 다루는 내용이다. 187324와 61700은 애니메이션으로 주인공이 모험을 헤쳐 나가며, 해적과 관련된 이야기이다.

본 논문에서 제안한 방법을 통해 사용자가 시청한 영화를 대상으로 이와 유사한 영화를 추천 할 수 있을 것으로 보인다.

4. 결론 및 향후연구

스마트 기기의 보유율이 꾸준히 증가하고 코로나 여파로 인해 스마트 기기 사용률도 함께 증가하고 있다. 특히 스마트 기기를 이용한 미디어 콘텐츠 이용율과 이용 시간이 크게 늘고 있으며, 주로 동영상제공서비스(OTT)를 이용하는 비율이 크게 증가했다. OTT는 콘텐츠 다양성, 요금제 적절성, 추천 시스템, N 스크린 서비스, 몰아보기 기능의 요소로 인해 이용 만족도와 지속 사용 의도가 결정된다. 그 중 추천 시스템의 영향이 가장 큰 것으로 연구되었다. 그러나 대부분의 추천 시스템은 사용자에게 메타 데이터를 강요하거나 화제성 높은 콘텐츠를 추천한다. 그렇기 때문에 콘텐츠 자체를 분석하여 사용자에게 추천하는 것은 매우 중요하다. 콘텐츠의 텍스트와 이미지를 함께 분석하기 위한 방법으로 조인트 임베딩이 주로 쓰인다. 하지만 대부분 이미지-텍스트 검색, 이미지 캡셔닝, Visual Question Answering(VQA)의 작업에 이용하며, 콘텐츠를 분석하는 연구는 미비하다. 따라서 본 논문에서는 텍스트와 이미지로 구성된 콘텐츠를 분석하기 위한 삼중항 손실 기반의 조인트 임베딩을 이용한 영상 콘텐츠 유사도 측정 방법을 제안한다. 영화를 도메인으로 하며, 텍스트는 줄거리, 이미지는 스틸컷과 포스터를 사용한다. 텍스트 임베딩에 KoBERT를 이용하고 이미지 특징 추출에 EfficientNet을 이용했으며, 이를 결합한 뒤 삼중항 손실을 이용하여 영화 콘텐츠의 거리를 학습했다. 그 결과 영화 콘텐츠 간 임베딩을 성공적으로 도출했으며, 유사함을 확인하기 위해 실제로 영화의 정보를 비교하여 타당함을 확인했다. 본 논문에서 제안한 방법을 통해 사용자가 시청한 콘텐츠를 대상으로 이와 유사한 콘텐츠를 추천할 때 활용 가능할 것으로 보인다.

향후 연구로 삼중항 손실을 수정 및 보완하여 새로운 삼중항 손실을 제안하고 우수한 임베딩 결과를 보이는 연구를 수행할 예정이다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1A2C2007091).

참 고 문 헌

[1] 신지형, 김윤화. “KISDI STAT REPORT 2020년 한국 미디어패널 조사결과 주요 내용”, 『정보통신정책연구원 ICT데이터사이언스연구본부』, 21-01호, 2021.

[2] 정용찬, 김윤화. “2020 방송매체 이용행태 조사“, 『방송통신위원회』, 2020.

[3] 정용국, 장위, “구독형OTT 서비스특성이이용자만족과 지속사용의도에미치는영향: 넷플릭스이용자를대상으로“, 한국콘텐츠학회논문지, 제20권, 제12호, pp. 123-135, 2020.

[4] Yan Gong, Georgina Cosma, Hui Fang, “On the Limitations of Visual-Semantic Embedding Networks for Image-to-Text Information Retrieval”, Imaging. vol.7 issue.8, July 2021.

[5] Gil Sadeh, Lior Fritz, Gabi Shalev, Eduard Oks, “Joint Visual-Textual Embedding for Multimodal Style Search”, arXiv preprint arXiv:1906.06620, Jun 2019.

[6] Liwei Wang, Yin Li, Svetlana Lazebnik, “Learning Deep Structure-Preserving Image-Text Embeddings”, CVPR 2016, Jun 2016.

[7]. Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurlio Ranzato, Tomas Mikolov, "DeViSE: A Deep Visual-Semantic Embedding Model", NIPS 2013, December 2013.

[8] Tomax Mikolov, Greg Corrado, Kai Chen, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space”, ICRL 2013, Juanuary 2013.

[9] Pengfei Liu, Xipeng Qiu, Xuanjing Huang, “Recurrent Neural Network for Text Classification with Multi-Task Learning”, IJCAI 2016, July 2016.

[10] Hasim Sak, Andrew Senior, Francoise Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling”, INTERSPEECH 2014, September 2014.

[11] 이상아, 장한솔, 백연미, 박수지, 신효필. “소규모 데이터 기반 한국어 버트 모델”, 『정보과학회논문지』, 제47권, 제7호, pp.682-692, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, CVPR 2016, June 2016.

[13] Mingxing Tan, Quoc V.Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”, PMLR 2019, June 2019.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, “Going deeper with convolutions”, CVPR 2015, June 2015.

[15] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, Sanja Fidler, “VSE++: Improving Visual-Semantic Embeddings with Hard Negatives”, BMVC 2018, September 2018.

공지사항

스마트미디어학회 조직위원회를 대표하여, 2021 추계학술대회를 보다 잘 준비하기 위한 협력에 감사드립니다. 논문 제출과 더불어 해당 논문의 분야에 대해 조사하고자 합니다.

논문이 수락된 경우, 학회 프로그램 세션에 분류될 트랙을 선택해주시기 바랍니다.

아울러 학문후속세대(학부생) 논문인 경우에는 아래의 칸에 추가 표시를 부탁합니다.

[ ] Smart Energy ICT

AMI(지능형 계량시스템), EMS(에너지관리시스템), BEMS(건물에너지관리시스템)
스마트홈, IoT, 스마트그리드, 마이크로그리드, 송변전자동화시스템, 배전자동화시스템
MDMS(계량데이터관리시스템), 전력거래시스템, EV, 분산전원, VPP(가상발전소)
ESS(에너지저장시스템), V2G(Vehicle to Grid)

[O] Smart Information

지능형컴퓨터, 클라우드컴퓨팅, 분산 및 병렬처리시스템, 인공지능, 영상처리
컴퓨터그래픽스, 음성처리, 멀티미디어, HCI, 빅데이터, 지능정보처리, 정보보호
모바일정보통신, 사물인터넷, 자동제어, 반도체, Microwave/Wireless, Optics

[ ] Information System

정보시스템 조직과 관리, e-비즈니스, ERP, CRM, SCM, 스마트워크, 소셜네트워크
IT아웃소싱, 프로젝트관리, 스마트라이프, 스마트 물류/금융/농업/교통/헬스케어
산업융합보안, 개인정보/의료정보/금융정보/산업기술보호, 스마트그리드, AMI

[ ] Contents & Services

융복합콘텐츠, 게임, 애니메이션, 웹/모바일, 스마트러닝, 문화디자인, 유니버셜디자인
UI/UX, 인터랙션 디자인, 디자인매니지먼트, 정보디자인, 디자인마케팅, 디자인방법론
디자인이론

[ ] Smart Media

미디어융합, 융복합 미디어, 디지털사이니지, 스토리텔링, 미디어콘텐츠와 기획, 창작,
전송유통, 마케팅

==========

[ ] 학문후속세대(학부생) 논문