

네이버 웹툰 썸네일로 저연령층 대상 콘텐츠 필터링

2022 년 06 월 01 일

서울과학기술대학교

빅데이터경영공학과

18102002 이현진

초록

웹툰의 인기는 꾸준히 증가해왔다. 특히 코로나 19로 인해 비대면 활동이 활발해지면서 웹툰을 즐기는 사용자의 수는 전반적으로 증가하였다. 이에 따라 국내 웹툰 시장 규모도 커지는 추세이며 유료 사용자 또한 증가하여 매출도 자연스럽게 늘어나고 있다. 증가하는 웹툰의 사용량에 따라 전 연령층이 접근 가능한 웹툰의 나이 제한 필터링 기능이 필요하다고 판단되었다. 본 연구는 로그인으로 인해 되어 구독자의 연령대를 파악할 수 있는 경우에 자극적인 장르를 필터링할 수 있도록 하는 데에 목적을 두었다.

웹툰 시장의 꾸준한 확대에도 불구하고 웹툰을 이용할 수 있는 플랫폼은 네이버 웹툰이 독점하고 있다고 볼 수 있다. 이에 웹툰 이용자들의 대부분이 사용하는 네이버 웹툰을 대상으로 연구를 진행했다. 저연령층을 대상으로 웹툰을 필터링하기 위해서는 필터링을 판단할 대상이 필요하다. 웹툰을 선택하는 중요한 요소이면서 대표하는 것으로 썸네일을 말할 수 있다. 이에 따라 썸네일을 해당 웹툰을 대표하는 대상으로 두고 자극적인 장르와 그렇지 않은 장르를 구분해보고자 한다.

자극적인 장르와 그렇지 않은 장르를 예측하기에 앞서 네이버 웹툰의 장르 총 10가지 중 액션, 스릴러, 무협/사극 장르를 자극적인 라벨링으로 두고, 그렇지 않은 장르를 자극적이지 않은 라벨링으로 구분해주었다. 이후 이미지 분류 모델을 다양하게 시도하여 학습시킨다. CNN, VGGNet, ResNet, EfficientNet의 여러 가지 모델들을 이용하여 가장 성능이 좋은 모델을 탐색해보고 성능을 높이기 위한 하이퍼파라미터 튜닝 결과를 제시한다. 성능 측정을 위한 방법으로는 정확도를 기반으로 판단하며 정확도를 높이는 방법으로 모델 선택과 하이퍼파라미터의 값이 결정된다.

본 연구의 기대효과로는 저연령층을 대상으로 하는 자극적인 장르의 웹툰 필터링을 넘어서 사용자 기반 맞춤형 웹툰 추천에 활용할 수 있다는 점이다. 이 과정에서 썸네일 기반 웹툰의 분류 모델은 매우 유용할 것이다. 썸네일 뿐만 아니라 줄거리도 함께 고려하여 그림체와 별개로 소재의 자극성을 판단하거나 사용자 맞춤형 추천에 활용할 수도 있다.

또한, 본 연구의 한계점 및 개선 방안으로는 장르를 자극적인과 그렇지 않은으로 분류하는 명확한 기준이 존재하지 않았다는 점이다. 그리고 네이버 웹툰에 국한되어 현저히 적은 훈련 데이터셋을 활용했다는 점에서 다양한 웹툰 플랫폼으로 훈련 대상을 확장하여 더 많은 데이터를 대상으로 학습할 필요성이 있다.

목 차

1. 서론	
A. 연구 배경 및 필요성 -----	1
B. 연구 목적 -----	2
C. 연구 구성 -----	3
2. 문헌 연구	
A. 이론적 배경 -----	3
B. 방법론적 배경 -----	5
3. 연구 방법	
A. 연구 프레임워크 -----	7
B. 데이터 수집 및 전처리 -----	7
C. 모델링 -----	8
4. 연구 결과	
A. 데이터 수집 및 전처리 -----	8
B. 모델링 -----	10
5. 결론	
A. 연구의 결론 -----	13
B. 연구의 한계점 및 추후연구 -----	14
6. 참고 문헌 -----	15

1. 서론

A. 연구 배경 및 필요성

닐슨코리아클릭에 따르면, 언택트 시대 다양한 엔터테인먼트 산업이 큰 성장세를 보이는 가운데 웹툰 시장 역시 꾸준한 성장세가 지속되어왔음에도 불구하고 여전히 웹툰 이용률에서 성장세를 보이고 있다. 게다가 글로벌 사업 확장과 드라마, 영화, 게임 등의 IP 확장까지 더하며 크게 주목받고 있다.



1) Nielsen-Koreanclick Android & iOS Mobile Behavioral Data (2019.07~2020.07)

2) 네이버 / 카카오 2분기 실적 발표 자료

*웹툰/웹소설 시장: 웹툰 / 웹소설 서비스를 주로 제공하며, 7월 기준 도달률이 0.5% 이상인 업체를 합산

닐슨코리아클릭, 「IP 무한 확장의 시대, 웹툰/웹소설 시장 왕좌의 주인은?»

▲ 2019년도 7월과 2020년도 7월을 비교했을 때 총 이용 시간이 24% 성장

이에 웹툰 콘텐츠의 장점을 살펴보면, 첫째, 경제성이다. 소재, 장르, 분량의 제약이 덜한 것은 물론이고 창작자가 스토리 연재 중 직접 피드백을 받으며 빠르게 스토리를 수정 및 확인할 수 있어 독자들의 요구사항을 즉시 반영하여 소비가 잘 되는 쪽으로 점차 발전시킬 수 있고, 작품의 성공 가능성도 빠르게 검증해볼 수 있다. 제작 속도에서도 웹소설에 비해 20배 빠르다고 한다. 둘째, 시의성이다. 독자들의 피드백을 빠르게 반영 가능한 콘텐츠인 만큼 현재 사람들이 가장 관심있어하는 트렌드와 욕구를 잘 포착해내어 더욱 각광받을 수 있다.

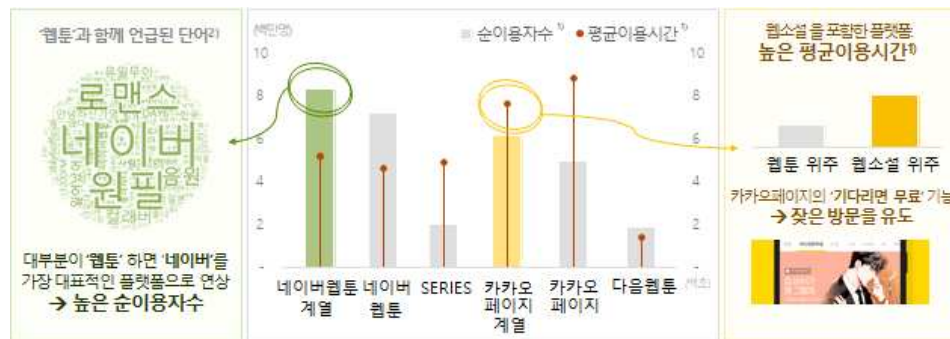
사용량이 꾸준히 증가하고 이용자층이 다양한 만큼 알고리즘 기반 사용자 맞춤형 콘텐츠를 제공할 필요가 있다. 현재 웹툰의 사용자 맞춤형 시스템은 유튜브나 넷플릭스 등의 영상 플랫폼과 비교했을 때 현저히 부족하다고 볼 수 있다. 특히, 저연령대를 대상으로 한 사용자 맞춤형 콘텐츠 제공이 필요하다. 인터넷 사용자층의 연령대가 낮아지면서 영상물은 어린이 프로그램을 필터링 하는 시스템을 기본적으로 갖추고 있다. 웹툰에는 영상물과는 다르게 글씨가 존재하여 영상물만큼 저연령층에게 자극적으로 다가가지는 않을 수 있지만, 게임

속의 그래픽으로 자극적인 상황을 연출하는 것처럼 웹툰 내에 표현된 그림들이 저연령층에게는 자극으로 다가갈 수 있다. 따라서 전 연령층이 접근 가능한 웹툰의 증가하는 사용량에 따라 웹툰의 나이 제한 필터링 기능이 필요하다고 판단되었다. 이를 위해 자극적인 콘텐츠를 분류해보고자 한다.

B. 연구 목적

본 연구는 저연령층을 대상으로 하는 자극적인 콘텐츠를 필터링하기 위해 썸네일을 이용하여 자극적인 장르를 예측하고자 한다.

닐슨코리아클릭에 따르면 소셜미디어 이용자들이 ‘웹툰’과 함께 언급한 단어 중 ‘네이버’가 가장 많았다고 한다. 이러한 연상 작용 덕분에 네이버 웹툰의 이용률이 가장 높다고 볼 수 있다. 이에 연구 대상 플랫폼으로는 “네이버 웹툰”을 대상으로 한다.



1) Nilesn-Koreanclick Android & iOS Mobile Behavioral Data (2020.07)

2) Social Media, Nielsen Buzzword, 2020.07.01~2020.07.31

*웹툰 위주 플랫폼: 웹툰 서비스를 주로 제공하며, 7월 기준 도달률이 0.5% 이상인 앱들을 합산 (네이버웹툰, 다음웹툰 등)

*웹소설 위주 플랫폼: 웹소설 서비스를 주로 제공하며 7월 기준 도달률이 0.5% 이상인 앱들을 합산 (카카오페이지, SERIES 등)

▲ 웹툰 시장의 1, 2위 사업자인 네이버와 카카오 비교

자극적인 콘텐츠를 판단하는 기준으로는 장르를 기반으로 한다. 네이버 웹툰에는 연재 방식에 따른 장르 구분, 콘텐츠에 따른 장르 구분이 있다. 이 중 콘텐츠 기반 자극적인 웹툰을 필터링하는 것이 목적이기 때문에 일상, 개그, 판타지, 액션, 드라마, 순정, 감성, 스릴러, 무협/사극, 스포츠 총 10가지의 장르를 액션, 스릴러, 무협/사극 장르를 자극적인 라벨링으로 두고 나머지 드라마, 순정, 감성, 일상, 개그, 판타지, 스포츠를 자극적이지 않은 라벨링으로 하여 진행하고자 한다.

해당 라벨링으로 장르를 분류한 기준은 썸네일이다. 썸네일은 매체와 장르의 특성을 반영한 고유의 디자인 패턴을 형성한 것으로 볼 수 있으며, 웹툰 내 다양한 그림체들과 분위기의 대표이다. 다른 7가지의 장르에 비해 액션, 스릴러, 무협/사극 장르의 경우 자극적인 그림들이 많이 나올 수밖에 없기 때문에 썸네일을 기준으로 자극적인 콘텐츠를 구분하기 위한 목적에 맞추어 해당 장르들을

자극적인 콘텐츠로 라벨링한다.

해당 연구를 진행하며 썸네일을 기반으로 자극적인 장르의 웹툰을 잘 필터링할 수 있는 성능 향상을 초점으로 탐색을 할 뿐만 아니라 이후 본 연구의 이용 확장 가능성까지 생각해보고자 한다.

C. 연구 구성

썸네일을 기반으로 장르를 예측하는 과정이기 때문에 웹툰의 썸네일과 장르에 대한 파악을 먼저 진행한다. 이후 이미지 처리를 위한 딥러닝 모델에 대한 검토 후 크롤링을 통해 얻은 네이버 웹툰 데이터들의 특성을 파악하여 적절한 전처리를 한다. 이후 다양한 딥러닝 모델들에 장르를 예측할 수 있는 학습을 진행하며 각 모델들을 대상으로 정확도 측면에서의 성능을 비교한다. 이후 해당 분류기 모델을 사용하여 사용자 맞춤형 시스템으로 확장할 수 있는 방향에 대한 탐구를 한다.

2. 문헌 연구

A. 이론적 배경

(1) 썸네일

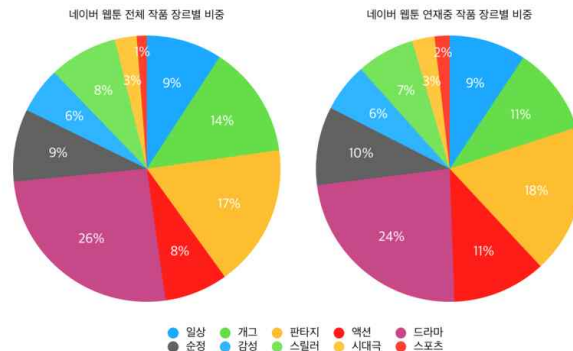
썸네일은 책 표지와도 같은 것으로 가장 먼저 독자의 시선을 끄는 이미지로서 작품의 매력을 함축한다. 웹툰 썸네일의 시초는 영화 포스터로, 영화는 웹툰에 앞서 가장 대중적인 스토리 매체였기에 유용한 참고자료이며, 영화 장르별 포스터 규칙을 웹툰 썸네일에도 비슷하게 적용할 수 있다. 그러나 영화와 달리 웹툰에는 유명 배우가 출연하지 않고 주요 장르와 소비층 역시 영화와는 다르며 모바일 세로 화면에 담긴다는 점도 큰 차이이다. 따라서 웹툰 썸네일은 매체와 장르의 특성을 반영한 고유의 디자인 패턴을 형성한다.



이에 웹툰 썸네일의 중요도를 테스트한 실험이다. 로맨스 판타지 <Finding Camellia>의 웹툰 썸네일을 여러 타입으로 만들어 A/B 테스트를 진행했고 유저 집단을 둘로 나누어 각기 다른 썸네일을 띄우고 반응을 살폈다고 한다. 결과치가 비슷할 거라고 예상했던 것과는 달리 최대 클릭률과 최소 클릭률의 차이가 8.38%p까지 벌어진 것으로 보아 웹툰에서 썸네일의 중요도가 크다는 것을 실험적으로 증명할 수 있다.

(2) 장르

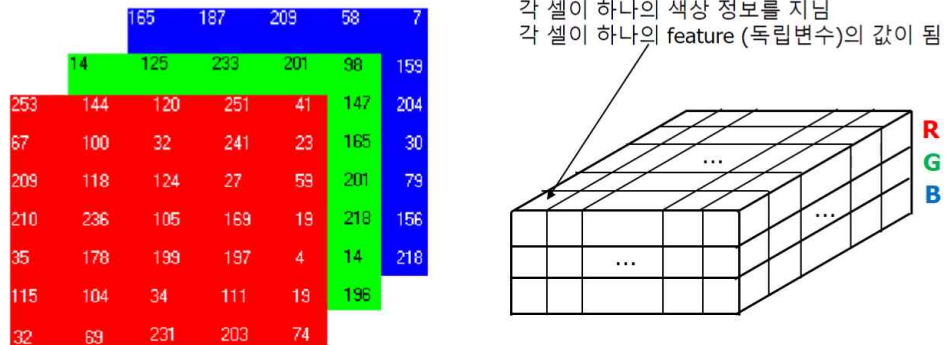
장르는 예술에서 작품을 구분할 때 이용되는 느슨한 분류 범위로 웹툰의 장르는 작가가 가장 잘 파악하겠지만, 장르에 구분에 대한 인식이 그리 높지 않거나, 작가도 선택하기 애매모호한 경우도 많이 존재한다. 따라서 장르를 판별할 수 있는 방법이 있다는 것은 매우 유용하다. 본 연구의 경우 장르를 기준으로 자극적인 웹툰 장르를 판별하는 만큼 자극적인 장르라고 분류할 수 있는 특정 장르들을 대상으로 저연령층 맞춤형 노출 여부 구분할 수도 있지만 해당 장르에 해당하는 모든 웹툰 콘텐츠를 노출시키지 않는 것보다는 이를 기반으로 학습한 모델을 바탕으로 자극적인 콘텐츠를 구분하는 것이 더 의미있다고 본다.



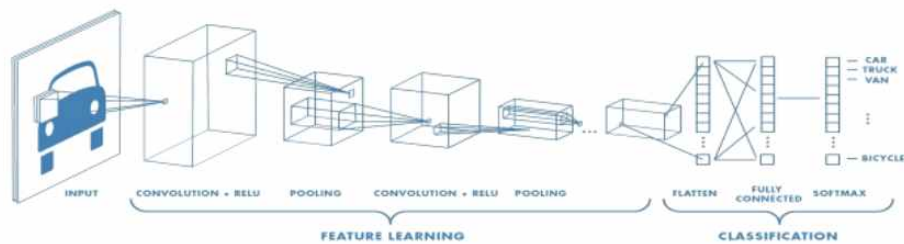
게다가 웹툰 장르의 구분은 각 매체마다, 작가마다 조금씩 다르다. 따라서 자극적이라고 판단되는 특정 장르를 지정하는 것보다는 썸네일의 분위기와 그림체들을 기반으로 학습한 모델을 적용하는 것이 앞으로 다른 웹툰 플랫폼이나 콘텐츠들에 적용하는 것에 더 의미있다고도 볼 수 있다. 그리고 독자층이 선호하는 장르가 있기 때문에 장르의 불균형이 존재할 수 밖에 없다. 장르의 불균형이 존재하는 데이터라는 점을 참고하여 의미있는 학습 결과가 나올 수 있는 방안을 찾아야 한다.

B. 방법론적 배경

(1) CNN(Convolutional Neural Network, 합성곱 신경망)



이미지 데이터는 여러 개의 픽셀로 구성되어 있고, 한 개의 픽셀은 3개의 색상정보를 저장하고 있다.

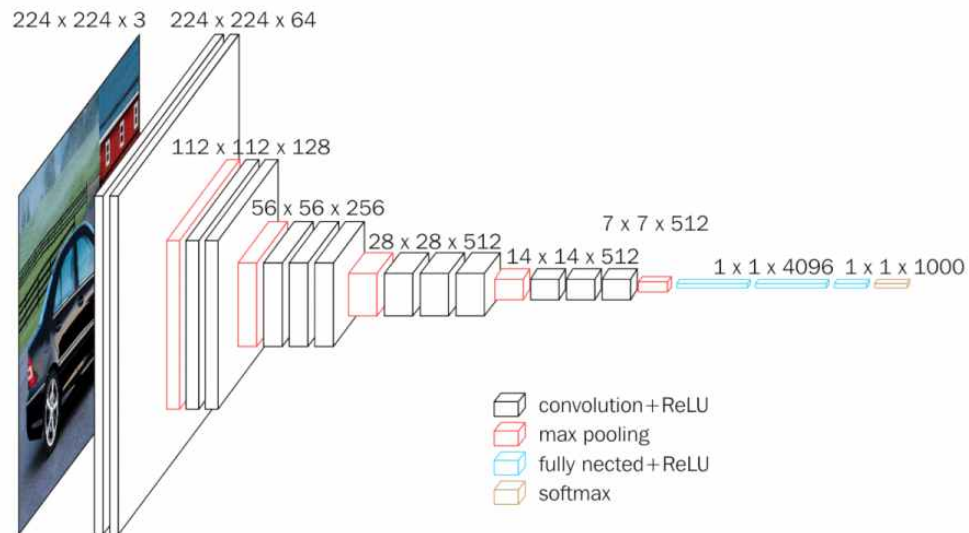


일반 신경망으로 이미지 데이터를 학습할 경우 3차원 이미지를 1차원으로 변환해서 입력하는 과정에서 정보 손실이 발생할 뿐만 아니라 학습 파라미터가 굉장히 많아 경제적으로 비효율적일 뿐만 아니라 과적합 가능성이 증가한다. 따라서 CNN은 필터를 이용해서 이미지의 각 셀들과 내적 연산(합성곱)을 수행하고 activation map을 생성하여 이미지의 각 부분의 정보를 추출하고자 한다. 이렇게 생성된 activation map을 또 하나의 이미지로 보고 다시 필터를 통해서 정보를 추출하는 과정을 반복한다. 그리고 최종적으로 1차원 데이터로 변환해서 softmax 함수를 출력층에 사용해서 이미지를 분류하는 구조이다.

이는 인간의 시신경 구조를 모방한 기술로, 사람이 여러 데이터를 보고 기억한 후 무엇인지 맞추는 것과 유사하다. 또한 2D인 이미지의 부분 정보를 추출하여 변환 없이 학습하기 때문에 이미지의 공간 정보를 유지한 채 학습을 하게 하는 모델이며 이미지 인식을 위한 패턴을 찾는 데 특히 유용하다. 특징맵을 생성하는 필터까지도 학습이 가능해 vision분야에서 성능이 우수하다. 이에 자율주행자동차, 얼굴인식과 같은 객체인식이나 computer vision이 필요한 분야에 많이 사용되고 있는 모델이다.

(2) VGGNet

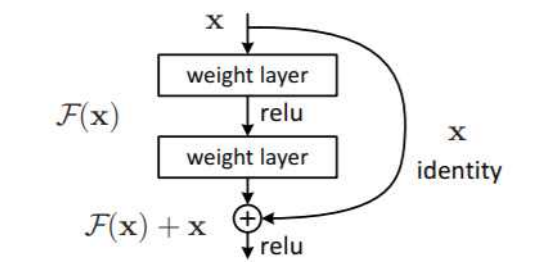
VGGNet은 네트워크의 깊이를 깊게 만드는 것이 성능에 어떤 영향을 미치는지를 확인하고자 컨볼루션 필터커널의 사이즈를 가장 작은 3×3 로 고정하고 진행한 실험으로부터 구성된 모델이다.



전체적인 구조는 3×3 필터를 stride=1로 설정해서 반복해서 적용하고 pooling layer를 제외하고 총 16개 층이어서 VGG-16이라고 부른다. pooling layer에는 2×2 로 max pooling을 stride=2로 해서 적용한다. VGG-16 과 VGG-19 중 주로 16 버전이 비교적 구조가 간단하고 사용이 편리해서 더 많이 사용된다. 이렇게 작은 필터를 사용함으로써 더 많은 ReLU함수를 사용할 수 있고 더 많은 비선형성을 확보할 수 있으며 더 좋은 성능을 가지게 된다.

(3) ResNet

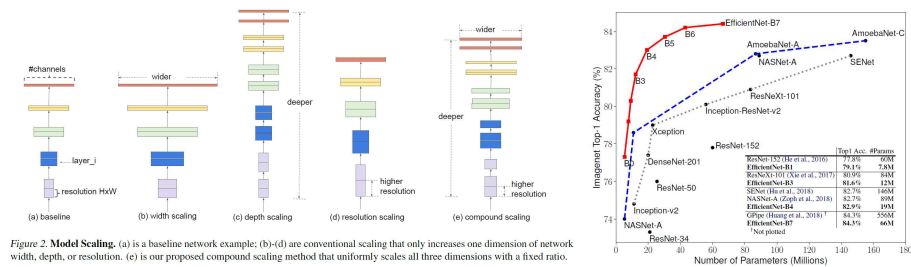
마이크로소프트에서 제안하고 2015년 이미지넷에서 우승한 모델로 당시 최초로 사람의 분류 성능을 뛰어넘으면서 인공지능에 대한 기대가 증가하게 되는 계기가 되었으며 최근까지도 가장 많이 사용되는 사전학습 모델이다.



일반적으로 신경망이 깊어지면 오히려 오차가 증가하고 경사 소실 (Gradient Vanisint)으로 인해 파라미터가 업데이트되지 않아 과적합된다. 이에 ResNet은 Skip Connection구조를 사용하여 입력값을 이용해서 출력값을 완전히 새로 계산하는 것이 아니라, 입력값과 원래 출력하고자 하는 값의 차이 (residual)만을 새롭게 학습하는 방식을 적용한다.

이를 통해 기존의 VGG 네트워크보다 더 깊지만 residual block을 활용해 복잡도와 성능은 더 개선한다. 그리고 구현이 간단하며, 학습 난이도가 매우 낮지며, 깊이가 깊어질수록 높은 정확도 향상을 보인다는 특징이 있다.

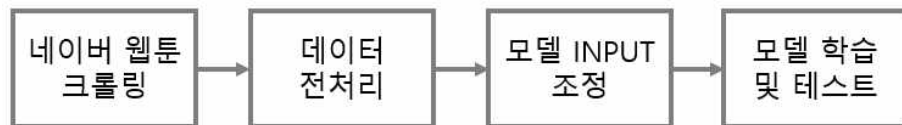
(4) EfficientNet



모델의 크기를 키움으로써 성능을 높이는 방향의 연구가 많이 이루어진다. 모델을 크게 만드는 방법으로는 network의 depth를 깊게 만드는 것, width가 넓을수록 미세한 정보가 많이 담아지기 때문에 channel width(filter 개수)를 늘리는 것, input image의 해상도를 올리는 것이 있다. EfficientNet은 이 3가지의 최적의 조합을 AutoML을 통해 찾은 모델로 2019년 구글에서 제안한 비교적 최신 기법이다.

3. 연구 방법

A. 연구 프레임워크



B. 데이터 수집 및 전처리

네이버 웹툰에 현재 연재중인 요일별 웹툰과 완결 웹툰을 대상으로 크롤링을 통해 썸네일, 장르 데이터를 얻는다. ‘일상’, ‘개그’, ‘판타지’, ‘액션’, ‘드라마’, ‘순정’, ‘감성’, ‘스릴러’, ‘무협/사극’, ‘스포츠’ 총 10가지의 장르 중 ‘액션’, ‘스릴러’, ‘무협/사극’은 “자극적인”으로,

‘일상’, ‘개그’, ‘판타지’, ‘드라마’, ‘순정’, ‘감성’, ‘스포츠’ 는 “자극적이지 않은” 으로 재분류한다. 그리고 재분류 결과 상대적으로 적은 개수의 장르가 포함되어 데이터의 개수가 적은 자극적인으로 분류된 콘텐츠의 수에 맞게 자극적이지 않은 장르에 해당하는 데이터를 핸드다운샘플링하며 train:valid:test = 8:1:1의 비율로 각 썸네일을 다운받아서 이미지를 저장한다. 이후 데이터 크기 조정, BATCH SIZE 결정, 이미지 augmentation 등 모델 INPUT DATA로 넣기 위한 과정을 진행한다.

C. 모델링

썸네일을 기반으로 웹툰을 1과 0으로 이진분류하는 모델을 학습 및 테스트한다. CNN, VGG16, ResNet, EfficientNet의 4가지 모델을 모두 진행하며 하이퍼파라미터 튜닝을 진행하며, 이후 각 모델의 정확도 및 loss값을 비교한다








4. 연구 결과

A. 데이터 수집 및 전처리




(1) 데이터 수집

네이버 웹툰에서 요일별로 연재중인 웹툰과 완결된 웹툰의 제목, 장르, 썸네일 주소 총 2515개를 크롤링한다.

요일별 전체 웹툰 인기순 업데이트순 조회순 별점순

월요일웹툰	화요일웹툰	수요일웹툰	목요일웹툰	금요일웹툰	토요일웹툰	일요일웹툰
						
창고속	김부장	확산귀환	연애혁명	의모자상주의	99강화나무동...	독립일기

완결 웹툰 인기순 업데이트순 조회순 별점순

		
유리와 유리하... 요연 ★★★★★ 9.97 전체보기	평범한 낙원 후드새 ★★★★★ 9.88 전체보기	이상형은 아님... 박윤영 / 은둥이 ★★★★★ 9.61 전체보기

level_0	index	title	genre	thumbnail
0	0	창고속	스토리, 액션	https://shared-comic.pstatic.net/thumb/webtoon/758037/thumbnail/thumbnail_IMAG06_794b0c1e-23aa-4c35-a335-b5d21b4bc2ab.jpg
1	1	소미다키크랑!	에피소드, 액션	https://shared-comic.pstatic.net/thumb/webtoon/783054/thumbnail/thumbnail_IMAG06_4856dbaf-26ce-4aee-bde5-87a0b1714022.jpg
2	2	뷰티풀 콘라리	스토리, 드라마	https://shared-comic.pstatic.net/thumb/webtoon/648419/thumbnail/thumbnail_IMAG06_44119122-6508-4293-9159-a99e2a0b0558.jpg
3	3	신의 힘	스토리, 판타지	https://shared-comic.pstatic.net/thumb/webtoon/183559/thumbnail/thumbnail_IMAG06_b1272b70-7eb4-4c1e-bc08-50b924e73be.jpg
4	4	웨스트지(상국의	스토리, 드라마	https://shared-comic.pstatic.net/thumb/webtoon/783052/thumbnail/thumbnail_IMAG06_f8e8c275-a327-48a9-9573-de64dad6d000.jpg
5	5	원드브레이커	스토리, 스포츠	https://shared-comic.pstatic.net/thumb/webtoon/602910/thumbnail/thumbnail_IMAG06_fd7fe8d8-e3f6-4553-88c2-46d861e49441.jpg
6	6	장씨제가 호위부사	스토리, 액션	https://shared-comic.pstatic.net/thumb/webtoon/728750/thumbnail/thumbnail_IMAG06_50aeb870-ea3e-47e4-bb38-a0d01e49ec8f.jpg
7	7	소녀의 세계	스토리, 드라마	https://shared-comic.pstatic.net/thumb/webtoon/654774/thumbnail/thumbnail_IMAG06_8ec78898-0b7f-44c3-9eb8-2ba9ab3b0790.jpg
8	8	백수세계	스토리, 드라마	https://shared-comic.pstatic.net/thumb/webtoon/733074/thumbnail/thumbnail_IMAG06_0d8a0c1c-30a2-4e72-94a9-59c272286c1e.jpg
9	9	팔이피플	스토리, 드라마	https://shared-comic.pstatic.net/thumb/webtoon/774883/thumbnail/thumbnail_IMAG06_58ecb055-dc10-43e8-bfa2-4a07f1c7f130.jpg

Show 10 per page

썸네일 주소에 해당하는 이미지를 나타내면 다음과 같다



▲ 첫 번째 데이터의 썸네일

(2) 데이터 전처리

먼저, 연재 방식과 콘텐츠 내용을 기준 총 두 가지로 표현된 장르를 콘텐츠 기반 장르 한 가지로 바꾸어준다

genre	genre
스토리, 액션	액션
에피소드, 액션	액션
스토리, 드라마	드라마
스토리, 판타지	판타지
스토리, 드라마	드라마
스토리, 스포츠	스포츠
스토리, 액션	액션
스토리, 드라마	드라마
스토리, 드라마	드라마
스토리, 드라마	드라마
스토리, 판타지	판타지
스토리, 판타지	판타지
스토리, 판타지	판타지
스토리, 판타지	판타지
스토리, 판타지	판타지
스토리, 드라마	드라마

이후 thumbnail에 있는 이미지 주소로부터 이미지를 전부 다운받아서 각 장르에 해당하는 폴더에 저장해준다. 이후 train, validation, test data로 각각 8:1:1의 비율로 옮기면서 액션, 스릴러, 무협/사극의 장르는 0 폴더에, 그 이외의 장르는 1 폴더로 옮긴다. 단, 0 폴더의 개수에 맞추어 1 폴더의 이미지 개수는 랜덤다운샘플링을 진행한다.

Action	test	
Daily	0	
Drama	1	
Fantasy	train	
Historical	0	
Joke	1	
Romance	valid	
Sentiment	0	
Sports	1	
Thriller		

total training 0 images: 444

total training 1 images: 469

total validation 0 images: 56

total validation 1 images: 59

total test 0 images: 57

total test 1 images: 59

마지막으로, CNN기반 Classification 이진분류 모델에 적용하기 위해 input 데이터를 조정한다. augmentation으로

rotation_range=40, width_shift_range=0.2, height_shift_range=0.2,
shear_range=0.2, zoom_range=0.2, horizontal_flip=True, fill_mode='nearest'

의 값을 주고, 그림의 크기는 (150, 150)으로 resize한 후 batch size는 20로 설정한다. batch size의 경우 여러 가지 진행해 본 결과 20일 때 가장 학습 속도 및 성능 측면에서 괜찮다고 판단되어 20으로 설정한다.

B. 모델링

공통적인 모델 파라미터는

Loss : binary_crossentropy, Metrics : accuracy, Steps per epoch : 45,

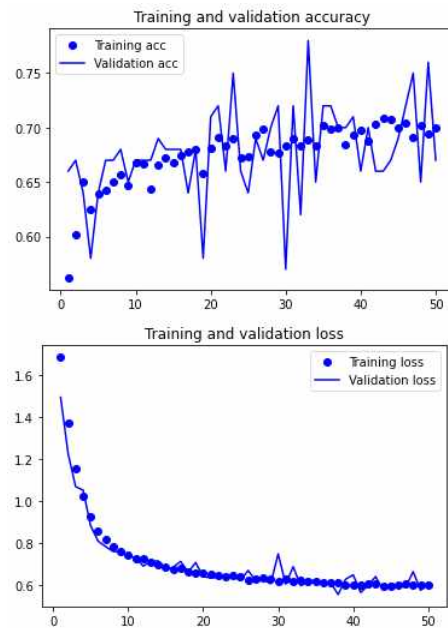
Validation_steps : 5, Epochs : 50

로 설정한다. epoch의 경우 50 이상으로 늘리면 과적합이 발생하여 모델 비교를 위해 50 으로 epoch를 고정한다.

(1) CNN

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 148, 148, 32)	896
max_pooling2d (MaxPooling2D)	(None, 74, 74, 32)	0
conv2d_1 (Conv2D)	(None, 72, 72, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 36, 36, 64)	0
conv2d_2 (Conv2D)	(None, 34, 34, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 17, 17, 128)	0
conv2d_3 (Conv2D)	(None, 15, 15, 128)	147584
max_pooling2d_3 (MaxPooling2D)	(None, 7, 7, 128)	0
flatten (Flatten)	(None, 6272)	0
dense (Dense)	(None, 128)	802944
dense_1 (Dense)	(None, 1)	129
Total params: 1,043,905		
Trainable params: 1,043,905		
Non-trainable params: 0		



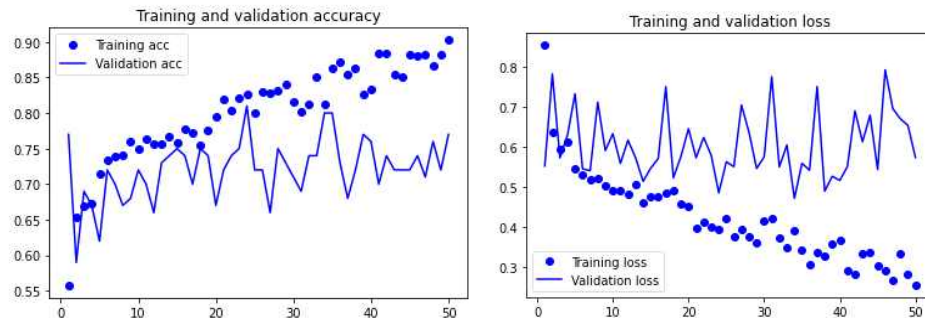
Optimizer : RMSprop(lr=1e-4)

Sequential 모델을 활용하여 직접 layer을 쌓았고 그 결과 loss측면에서는 안정적으로 유의미하고 accuracy 측면에서는 안정적이지는 않지만 점차 개선되는 추세를 보였다. 이를 test data에서 성능평가 했을 때에 loss값 0.58, accuracy값 0.72이다.

(2) VGG16

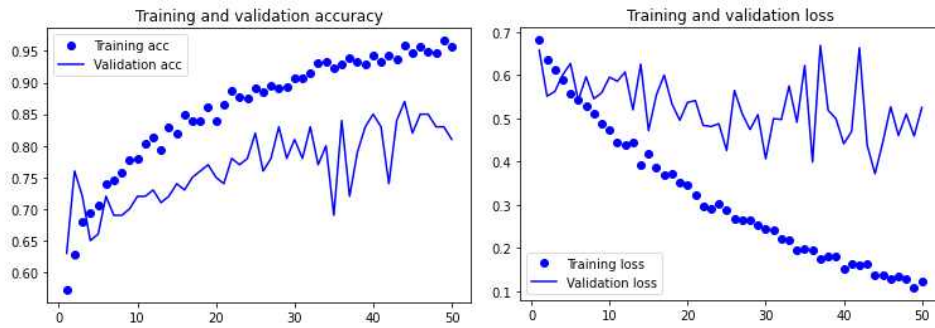
Model: "vgg16"					
Layer (type)	Output Shape	Param #			
input_1 (InputLayer)	[(None, 150, 150, 3)]	0			
block1_conv1 (Conv2D)	(None, 150, 150, 64)	1792			
block1_conv2 (Conv2D)	(None, 150, 150, 64)	36928			
block1_pool (MaxPooling2D)	(None, 75, 75, 64)	0			
block2_conv1 (Conv2D)	(None, 75, 75, 128)	73666			
block2_conv2 (Conv2D)	(None, 75, 75, 128)	147584			
block2_pool (MaxPooling2D)	(None, 37, 37, 128)	0			
block3_conv1 (Conv2D)	(None, 37, 37, 256)	295168			
block3_conv2 (Conv2D)	(None, 37, 37, 256)	590080			
block3_conv3 (Conv2D)	(None, 37, 37, 256)	590080			
block3_pool (MaxPooling2D)	(None, 18, 18, 256)	0			
block4_conv1 (Conv2D)	(None, 18, 18, 512)	1180160			
block4_conv2 (Conv2D)	(None, 18, 18, 512)	2359808			
block4_conv3 (Conv2D)	(None, 18, 18, 512)	2359808			
block4_pool (MaxPooling2D)	(None, 9, 9, 512)	0			
block5_conv1 (Conv2D)	(None, 9, 9, 512)	2359808			
block5_conv2 (Conv2D)	(None, 9, 9, 512)	2359808			
block5_conv3 (Conv2D)	(None, 9, 9, 512)	2359808			
block5_pool (MaxPooling2D)	(None, 4, 4, 512)	0			
Total params: 14,714,688					
Trainable params: 14,714,688					
Non-trainable params: 0					

Model: "sequential_1"					
Layer (type)	Output Shape	Param #			
vgg16 (Functional)	(None, 4, 4, 512)	14714688			
flatten_1 (Flatten)	(None, 8192)	0			
dense_2 (Dense)	(None, 256)	2097408			
dense_3 (Dense)	(None, 1)	257			
Total params: 16,812,353					
Trainable params: 2,097,665					
Non-trainable params: 14,714,688					



Optimizer : Adam(lr=0.001)

VGG16의 imagenet으로 사전학습 된 모델을 불러와서 사용하고 마지막 Fully connected layer만 직접 쌓아준 후 직접 쌓은 부분만 가중치 업데이트가 가능하도록 한다. 이 결과 test data에서 성능평가 했을 때에 loss값 0.92, accuracy값 0.66이며, 그래프도 매우 불안정하고 test data의 성능도 매우 좋지 않음을 알 수 있다. 이에 가중치 업데이트가 가능한 층의 범위를 넓혀 사전 학습된 모델까지 학습하도록 해 보기로 했다. 사전학습된 모델의 마지막 convolution block까지 가중치 업데이트가 가능하도록 모델의 튜닝 범위를 변경하여 한번 더 진행해본다.



Optimizer : RMSprop(lr=1e-4)

그 결과 test data에서 성능평가 했을 때에 loss값 1.13, accuracy값 0.69라는 결과를 얻을 수 있었으며, 정확도 기준으로 판단했을 때 아주 약간 성능이 개선된 것으로 보이지만, loss값은 심하기 불안정하며 과적합되는 모습을 볼 수 있다. 단, 정확도 기준 모델 판별이므로 VGGNet모델은 해당 finetuning방법을 택한다.

(3) ResNet

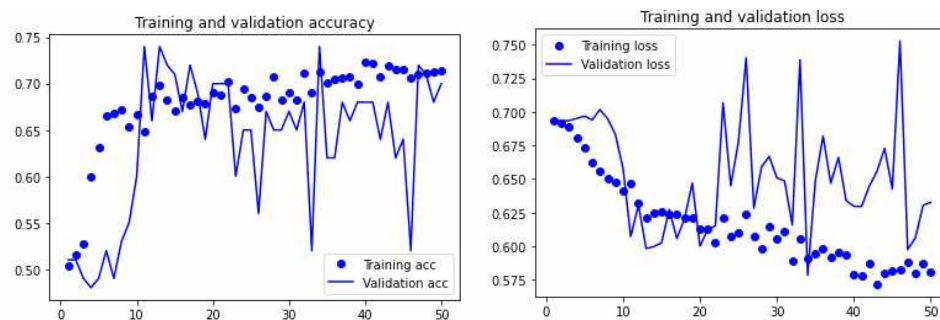
Model: "sequential_5"

Layer (type)	Output Shape	Param #
resnet50 (Functional)	(None, 2)	23591810
dense_11 (Dense)	(None, 512)	1536
dropout_2 (Dropout)	(None, 512)	0
dense_12 (Dense)	(None, 512)	262656
dropout_3 (Dropout)	(None, 512)	0
dense_13 (Dense)	(None, 1)	513

=====

Total params: 23,856,515
Trainable params: 23,803,395
Non-trainable params: 53,120

=====



Optimizer : RMSprop(lr=2e-5)

tensorflow의 ResNet50 모델을 사용했으며, test data에서 성능평가 했을 때에

loss값 0.67, accuracy값 0.66이며 훈련 결과 매우 불안정하며 성능도 좋지 않은 것으로 보인다.

(4) EfficientNet

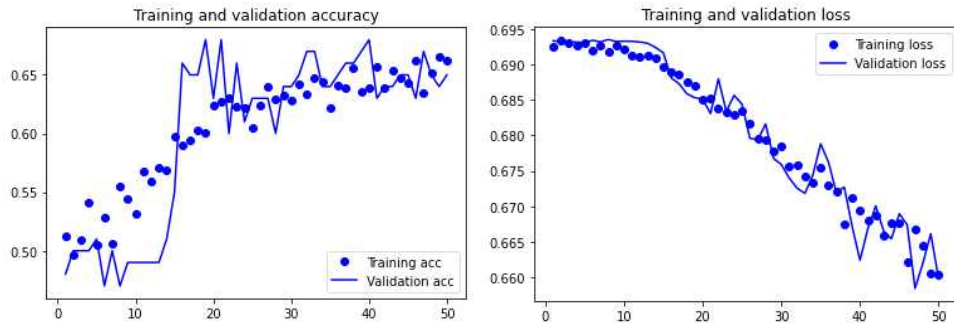
Model: "sequential_6"

Layer (type)	Output Shape	Param #
efficientnetb0 (Functional)	(None, 2)	4052133
flatten_4 (Flatten)	(None, 2)	0
dense_14 (Dense)	(None, 256)	768
dense_15 (Dense)	(None, 1)	257

=====

Total params: 4,053,158
Trainable params: 4,011,135
Non-trainable params: 42,023

=====



Optimizer : Adam(lr=1e-5)

tensorflow의 EfficientNetB0모델을 사용했으며, test data에서 성능평가 했을 때에 loss값 0.65, accuracy값 0.69이며 훈련 결과 비교적 과적합 없이 잘 훈련된 것에 비해 test data의 성능이 높지는 않다.

5. 결론

A. 연구의 결론

	Train_loss	Train_acc	Valid_loss	Valid_acc	Test_loss	Test_acc
CNN	0.5958	0.6999	0.6133	0.6700	0.5839	0.7155
VGGNet	0.1214	0.9574	0.5251	0.8100	1.1300	0.6897
ResNet	0.5803	0.7144	0.6237	0.7000	0.6743	0.6638
EfficientNetB0	0.6604	0.6618	0.6598	0.6500	0.6498	0.6897

Train, Validation 지표는 마지막 epoch를 기준으로 작성된 수치이다. 표를 통해 EfficientNet 모델이 훈련 그래프가 가장 안정적이었지만 전체적인 성능은 CNN이 test data의 정확도 측면에서 가장 좋다는 사실을 알 수 있다.

전반적으로 loss값이 매우 크게 측정된다. 이는 0과 1의 클래스를 예측하는 정도가 명확하지 않거나 잘못 예측할 때 경우에 강한 확률값으로 잘못 예측하면 정확도와는 별개로 loss값이 클 수 있다. 확실한 결론을 위해서는 정확도 개선의 방안도 필요하지만 loss값을 줄이는 방안이 먼저 필요하다고 생각되고, 이 방안으로는 적은 데이터셋이 가장 큰 원인이라고 생각되어 훨씬 많은 웹툰 썸네일이 존재한다면 개선 가능한 것으로 판단된다. 마찬가지로 적은 데이터셋의 문제가 해결된다면, 모델 구조의 단순성으로 인해 조금이라도 test data의 정확도 측면에서 가장 좋은 성능을 보였던 CNN보다는 전체적으로 안정적인 결과를 도출한 EfficientNet의 결론이 가장 좋을 것으로 예상된다.

이를 바탕으로, CNN기반의 모델을 통해 자극적인 장르와 그렇지 않은 장르를 구분하는 것의 방법론들을 살펴볼 수 있었다.

B. 연구의 한계점 및 추후연구

본 연구의 가장 큰 한계점은 현저히 적은 데이터셋으로 인해 모델의 훈련 결과들이 불안정하다는 것이다. 이는 추후 다른 플랫폼들의 데이터를 대상으로 진행하며 개선 가능할 것으로 예상된다. 또한, 해당 연구는 저연령층을 위한 자극적인 장르와 그렇지 않은 장르에 대한 분류 모델 탐색이었지만, 이를 사용자 맞춤형 시스템 제공으로 활용하기 위해서는 다양한 장르들을 대상으로 할 수도 있다고 생각하고 이 경우 장르들의 군집화가 필요하다고 판단된다. 그 이유로는 심한 장르의 불균형 문제와 비슷한 장르의 존재 때문이고 이를 군집화하기 위한 명확한 기준이 필요하다. 마찬가지로 해당 연구의 ‘자극적인’과 ‘자극적이지 않은’을 나눈 방법이 주관적이었던 것에 아쉬움이 있어서 추후 이를 분류하는 것에 대한 명확한 기준 성립이 필요해 보인다.

이 연구를 토대로 썸네일 뿐만 아니라 줄거리도 고려한 사용자 맞춤형 시스템 구현으로 확장할 수 있다. 또한 웹툰의 썸네일의 대상으로 저연령층 대상 콘텐츠 필터링을 한 만큼 웹툰 뿐만 아니라 어린이용 플랫폼의 필터링에도 이용 가능하다.

6. 참고 문헌

[닐슨코리아 클릭]

http://www.koreanclick.com/insights/newsletter_view.html?code=topic&id=586

[이론적 배경]

<https://ridicorp.com/story/pr-webtoon-thumbnail-app-design/>

<https://m.blog.naver.com/enterani/221961425299>

[방법론적 배경]

<https://yeong-jin-data-blog.tistory.com/entry/%EC%82%AC%EC%A0%84%ED%95%99%EC%8A%B5-%EB%AA%A8%EB%8D%B8-CNN>

<https://rubber-tree.tistory.com/entry/%EB%94%A5%EB%9F%AC%EB%8B%9D-%EB%AA%A8%EB%8D%B8-CNN-Convolutional-Neural-Network-%EC%84%A4%EB%AA%85>

<https://daechu.tistory.com/10>

<https://bskyvision.com/504>

<https://daeun-computer-uneasy.tistory.com/28>

<https://greeksharifa.github.io/computer%20vision/2022/03/01/EfficientNet/>

*** 프로젝트 코드는 해당 colab 링크로 들어가주세요!***

https://colab.research.google.com/drive/16oFFUIDgrSMuwsBpEWSJ8_Vlgi8Aa3q6?usp=sharing