creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

[UCI]I804:24011-200000265867

2017년 2월

석사학위 논문

CNN을 이용한 SNS 사용자 관심사 카테고리 분류 및 팔로잉 추천 방법

조선대학교 산업기술융합대학원

소프트웨어융합공학과

홍 택 은

조선대학교
CHOSUN UNIVERSITY

2017년 2월 석사학위논문

CNN을 이용한 사용자 관심사 카테고리 분류 및 팔로잉 추천 방법

홍 택 은

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

CNN을 이용한 SNS 사용자 관심사 카테고리 분류 및 팔로잉 추천 방법

Interest Category Classification of SNS User and Following Recommendation Method using CNN

2017년 2월 24일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

홍 택 은

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

CNN을 이용한 SNS 사용자 관심사 카테고리 분류 및 팔로잉 추천 방법

지도교수 신 주 현

이 논문을 공학석사학위신청 논문으로 제출함.

2016년 10월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

홍 택 은

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

홍택은의 석사학위논문을 인준함

위원장 조선대학교 교수 반 성 범 (인)

위 원 조선대학교 교수 곽 근 창 (인)

위 원 조선대학교 교수 신 주 현 (인)

2016년 11월

조선대학교 산업기술융합대학원

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

목 차

ABSTRACT

Ⅰ. 서론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 연구 내용 및 구성 ····· 2

Ⅱ. 관련 연구 ····· 3
A. 관심사 분류 관련 기존 연구 ····· 3
B. Convolutional Neural Network ····· 6
1. 이미지 분류 ····· 9
2. 텍스트 분류 ····· 11
C. TensorFlow ····· 13
D. DMOZ ····· 15

Ⅲ. 관심사 카테고리 분류를 이용한 팔로잉 추천 방법 ····· 17
A. 시스템 구성도 ····· 17
B. 정보제공자 데이터베이스 구축 ····· 18
1. 관심사 카테고리 정의 ····· 19
2. CNN을 이용한 관심사 카테고리 분류 ····· 20
a. 이미지에서 관심사 카테고리 분류 ····· 20
b. 텍스트에서 관심사 카테고리 분류 ····· 25
3. 관심사 카테고리를 이용한 데이터베이스 구축 ····· 29
C. 사용자 팔로잉 추천 ····· 32
1. 사용자 관심사 카테고리 분류 ····· 32
2. 거리 측정을 통한 팔로잉 추천 ····· 35

Ⅳ. 실험 및 결과 ····· 39
A. 실험 환경 ····· 39
B. 데이터 셋 ····· 40
C. 실험 평가 및 분석 ····· 41
1. 실험 평가 방법 ····· 41

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 실험 결과 분석 ····· 42

Ⅴ. 결론 및 제언 ····· 47

참고문헌 ····· 48

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

표 목 차

[표 2-1] TensorFlow 용어 ····· 14
[표 3-1] 관심사 카테고리 ····· 19
[표 3-2] 정보제공자 데이터베이스 ····· 30
[표 3-3] 사용자 관심사 분류 테이블 ····· 33
[표 3-4] 관심사 카테고리가 Arts인 정보제공자와 사용자의 이미지 예 ····· 36
[표 3-5] 관심사 카테고리가 Arts인 정보제공자와 사용자의 텍스트 예 ····· 36
[표 3-6] 정보제공자와 사용자간 거리 측정 테이블 ····· 37
[표 4-1] 실험 환경 ····· 39
[표 4-2] 페이스북에서 텍스트와 이미지 수집 ····· 40
[표 4-3] 전체 데이터 셋 ····· 41
[표 4-4] 정보제공자 관심사 카테고리 분류 성능 ····· 43
[표 4-5] 사용자 관심사 카테고리 분류 성능 ····· 43

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

그 림 목 차

[그림 2-1] 스마트 기기 사용자 증가율 ····· 3
[그림 2-2] SNS 사용자 증가율 ····· 3
[그림 2-3] Convolutional Neuronal Network ····· 6
[그림 2-4] Convolution layer에서 특징 값 추출 ····· 7
[그림 2-5] filter의 이동 모습 ····· 8
[그림 2-6] channel에서 Max-Pooling 이용의 예 ····· 8
[그림 2-7] ILSVRC의 우승팀, top-5 error rate, layer의 개수 ····· 9
[그림 2-8] 텍스트 분류를 위한 CNN 구성 ····· 11
[그림 2-9] 텐서 플로우 그래프의 예 ····· 13
[그림 2-10] TensorBoard의 예 ····· 14
[그림 2-11] DMOZ 카테고리 ····· 15
[그림 2-12] DMOZ 카테고리의 형태 ····· 16
[그림 3-1] 전체 시스템 구성도 ····· 17
[그림 3-2] 정보제공자 데이터베이스 구성도 ····· 18
[그림 3-3] 학습을 위해 수집한 이미지의 예 ····· 21
[그림 3-4] Inception-v3의 기본 구조 ····· 21
[그림 3-5] 이미지 분류를 위한 CNN 모델의 플로우 차트 ····· 22
[그림 3-6] 이미지 분류를 위한 CNN 모델의 학습 결과 ····· 23
[그림 3-7] 정보제공자 이미지에서 관심사 분류 결과의 예 ····· 24
[그림 3-8] 학습을 위해 수집 후 전처리를 거친 텍스트의 예 ····· 25
[그림 3-9] 텍스트 분류를 위한 CNN 모델의 플로우 차트 ····· 26
[그림 3-10] 텍스트 분류를 위한 CNN 모델의 학습 내용 ····· 27
[그림 3-11] 정보제공자 텍스트에서 관심사 분류 결과의 예 ····· 28
[그림 3-12] 정보제공자와 사용자 사이의 거리 유사도 측정 결과 ····· 38
[그림 4-1] 정보제공자 카테고리 분류 성능 그래프 ····· 44
[그림 4-2] 사용자 카테고리 분류 성능 그래프 ····· 44
[그림 4-3] 기존 방버과 제안하는 방법에 대한 성능 비교 ····· 45

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

Interest Category Classification of SNS User and Following Recommendation Method using CNN

Taekeun Hong
Advisor : Prof. JuHyun Shin, Ph.D
Department of SoftWare Convergence Engineering
Graduate School of Chosun University

With the development of various smart devices, an increasing number of people are using SNS, which enables communication and sharing of information regardless of distance and place. Users who used to focus on the functions of forming relationship and communication are now using SNS for the function of sharing information. The existing way of recommending following was mostly based on relationship so there is the limitation of recommending following for users who use SNS for information. Also, as studies using SNS postings mostly suggest ways of categorizing interests and recommending following based on images and texts, it is hard to understand user‘s intention exactly.

So this paper suggests the way of recommending information-centered following that finds out user’s interests by using images and texts and categorizing interests of users. This study secured objectivity by defining DOMZ, which is established as ODP(Open Directory Project), as the standard for category of interest, and by using CNN(Convolutional Neural Network), one of the machine learning methods, it studied images and text and divided interests into categories. It defined users who post many postings as information providers, established information database and categorized users

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

with the same interests. Then it measured the similarity of the distance between information providers and users and recommended the information provider with the most similar interests for following.

As a result of categorizing interests of information providers and of users using the method suggested in this paper, the rate of accuracy was 80% and 79.8% respectively, and the overall rate of accuracy was 79.93%, which means a good categorization overall. Categories of interest with low rate of accuracy are thought to be because the meanings were similar and similar data was collected when images and texts were searched. There seems to be a slight correlation between categories. Therefore, it seems that there need to be studies on the selection of categories and standards for making SNS categories, and also studies on establishing proper CNN models to learn and categorize texts that are made up of special characters and emoticons due to the nature of texts of SNS postings.

What can be suggested is to divide user's interests into 16 categories then recommend following in order to recommend more proper following for information-focused SNS users, and by using images and texts together, it was possible to accurately understand the intention of users. Through the way suggested by this paper, it is likely that it will be possible to recommend following that is suitable for SNS users, and categorization of interests will be used in customized services, which recommend contents and information suitable for user's interests, and SNS marketing services.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구 배경 및 목적

SNS(Social Network Service)는 거리, 공간의 제약 없이 사용자 간 실시간으로 의사소통, 관계 형성, 정보 공유를 할 수 있는 사이버 공간이다. 다양한 스마트 디바이스의 시장 규모가 급속히 증가하면서 스마트 디바이스를 보유하고 있는 인구도 함께 증가하였으며, SNS를 즐기는 사용자도 증가하였다[1,2]. 현재 출시된 SNS의 종류를 관계 형성 중심과 정보공유 중심으로 구분할 수 있다. 관계 형성 중심의 SNS의 종류에는 페이스북(Facebook), 트위터(Twitter) 등이 있으며, 주로 텍스트 위주로 이용할 수 있도록 서비스를 제공한다. 정보 중심의 SNS의 종류에는 인스타그램(Instagram), 텀블러(Tumblr), 플리커(Flickr) 등이 있으며, 주로 이미지 위주로 이용할 수 있도록 서비스를 제공한다[3]. SNS가 등장한지 얼마 되지 않았을 때는 사용자들이 의사소통, 관계 형성을 목적으로 서비스를 이용했다. 하지만 실생활에서 느끼는 인간관계에 대한 싫증과 지루함을 SNS를 이용하면서 느끼게 되면서 정보공유의 기능으로 활용하는 사용자가 늘고 있다[4]. 사용자는 SNS를 통해서 자신의 관심사에 해당하는 정보를 계정에 게시하고 관리하며, 관심사를 표출하고 비슷한 관심사를 갖는 사용자의 계정을 팔로잉함으로써 정보를 습득할 수 있다[2]. SNS에서 팔로잉 추천을 제공하고 있으나 대부분 관계에 중점을 둔 팔로잉 추천을 제공하고 있기 때문에 정보 중심의 팔로잉 추천에는 적합하지 않다. SNS에서 습득할 수 있는 정보를 이용하여 사용자의 특성이나 관심사 등을 판별하고 사용자에게 맞춤형 서비스를 제공하기 위한 연구가 활발하게 이루어지고 있으나 대부분 텍스트와 이미지를 각각 이용하거나 사용자에 대한 메타데이터를 이용하는 경우가 대부분이다.

따라서 본 논문에서는 SNS 게시글에 존재하는 텍스트와 이미지를 함께 이용하여 관심사 카테고리를 기준으로 사용자의 관심사를 분류하고 같은 관심사를 갖는 다른 사용자를 팔로잉 추천하고자 한다. 게시글에서 텍스트와 이미지를 모두 사용하기 때문에 각각 이용하는 경우보다 더 많은 정보를 활용하며, 기계학습을 이용하여 관심사 카테고리를 기준으로 하는 사용자의 관심사를 정확히 분류할 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 연구 내용 및 구성

본 연구의 주요 내용은 SNS 게시물을 이용하여 사용자의 관심사를 카테고리 기반으로 분류하고 팔로잉을 추천하는 방법이며, 이를 제안하기 위해 다음과 같은 구성으로 작성되었다.

서론에 이어 2장 관련 연구에서는 본 연구의 이론적인 배경이 되는 CNN(Convolutional Neural Network)을 비롯한 관심사 분류의 기존 연구들에 대해 살펴본다.

3장에서는 본 논문에서 제안하는 관심사 카테고리 분류를 이용한 팔로잉 추천 방법에 관해 기술한다. CNN으로 텍스트와 이미지의 관심사 카테고리를 분류하여 정보제공자 데이터베이스를 구축하고 사용자에게 팔로잉을 추천하는 방법에 대해 제안한다.

4장에서는 실험환경과 실험에 사용한 각종 데이터 셋에 관해 설명하고, 실험을 통한 비교 및 성능을 평가한다.

마지막으로 5장에서는 결론과 향후 연구에 관해 서술하며 마무리한다.

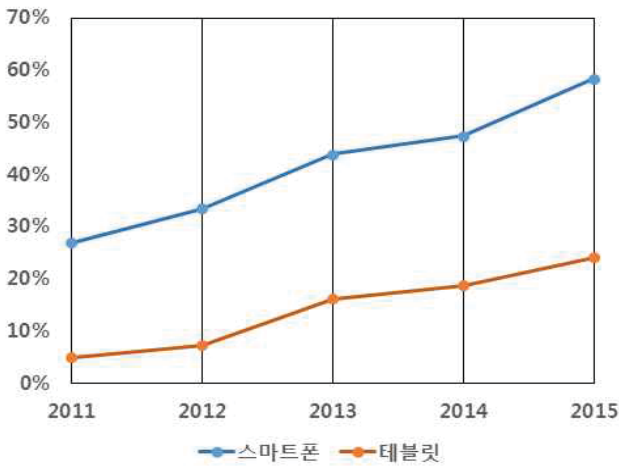
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

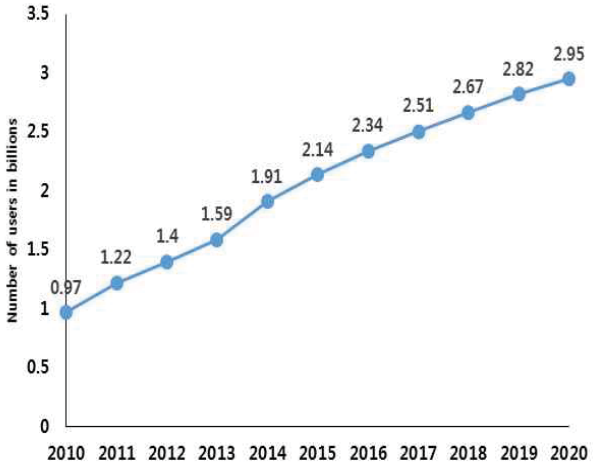
Ⅱ. 관련 연구

A. 관심사 분류 관련 기존 연구

스마트 기기의 발전과 보급률의 증가에 따라 스마트 기기를 이용하여 간편하게 즐길 수 있는 SNS 사용자 또한 증가하고 있다. [그림 2-1]은 스마트 폰과 테블릿의 보급률을 나타내는 그래프이며 [그림 2-2]는 소셜 미디어 사용자를 나타내는 그래프이다.



[그림 2-1] 스마트 기기 사용자 증가율



[그림 2-2] SNS 사용자 증가율

[그림 2-1]를 통해 2011년부터 2015년까지 스마트 기기를 사용하는 인구가 증가하고 있다는 것을 알 수 있으며, [그림 2-2]를 통해 SNS 사용자가 증가하고 있다는 것을 알 수 있으며[9,10], SNS 사용자가 증가함에 따라 게시되는 글 또한 증가하고 있다. 이를 이용한 다양한 연구가 활발하게 진행되고 있으며, 관심사 분류는 SNS를 이용한 연구 중 하나이다.

관심사란 관심을 끄는 일 또는 관심거리를 말하며, 관심사 분류는 사용자 맞춤형 서비스를 제공하기 위한 기반 연구가 되기 때문에 사용자의 정확한 관심사를 분류하는 것은 매우 중요하다. 사용자 맞춤형 서비스를 제공하기 위해 사용자가 게시글을 통해 생각이나 의견을 표출할 수 있는 서비스인 블로그, SNS 등을 주로 이용하

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

며, 분석하고자 하는 데이터의 종류에 따라 관심사를 분류하는 방법이 달라진다. 다양하고 방대한 SNS 게시글을 이용한 사용자 맞춤형 서비스는 사용자의 개별적인 특성이나 관심사를 분석하고 개인의 선호도에 알맞은 콘텐츠를 추천하기 위한 기법에 관한 연구이다[5,6,7,8,9,10].

[5]에서는 사용자의 프로 파일과 SNS 게시글을 이용하여 최근 관심사를 파악하고 뉴스 카테고리를 기반으로 뉴스 콘텐츠를 분석하여 사용자의 관심사와 뉴스 콘텐츠를 유사도 분석하여 적합한 콘텐츠를 추천하는 방법을 제안했다.

[6]에서는 SNS에 적합한 요소와 사용자 평가 이력을 이용하며, 협업 필터링을 적용하여 사용자 개개인의 취향을 명시적으로 반영할 수 있는 개인화 추천기법에 관한 연구를 했다.

[7]에서는 사용자가 작성한 글과 친구관계인 사용자와 주고받은 내용을 기반으로 카테고리화 시킴으로써 사용자의 관심사를 자동으로 분류 및 추천하는 방법을 제안했다.

[8]에서는 SNS에서 영화관련 게시글을 수집하여 Thayer 모델을 기반으로 감정 모델 분류 기준을 세우고 대중이 느끼는 감정을 분석한 후 유사도를 측정하는 연구를 했다.

[9]에서는 SNS 사용자 게시글의 텍스트에 LDA(Latent Dirichlet Allocation) 알고리즘을 적용하여 단어의 등장 확률분포를 통해 Topic을 추출하고 사용자의 관심사를 분류하는 방법을 제안했다.

[10]에서는 SNS 게시글의 텍스트와 이미지를 함께 사용하여 사용자의 관심사를 분류했다. 텍스트의 빈도수를 기반으로 관심사를 분류하고 태깅되지 않은 CNN을 사용하여 이미지를 분류한 후 텍스트와 이미지의 관심사를 비교하여 최종적으로 사용자의 관심사를 분류하는 방법을 제안했다.

앞서 설명했듯이 SNS 게시글을 이용하여 관심사를 분류하고 사용자에게 적합한 내용을 추천하거나 감정을 분석하는 다양한 관심사 기반의 연구가 진행되고 있다. SNS 게시글에는 텍스트뿐만 아니라 이미지도 존재하지만, 기존의 연구들은 대부분 텍스트만을 이용하여 관심사 분류를 했다. 이런 경우 게시글의 모든 데이터를 이용하지 않고 일부만 사용하기 때문에 사용자의 관심사나 의도를 정확하게 파악하는데 어려움이 존재하며, 텍스트와 이미지를 함께 사용하더라도 분류하고자 하는 관심사에 대한 기준이 존재하지 않으므로 정확한 관심사 분류가 쉽지 않다.

따라서 본 연구에서는 SNS 게시글에서 텍스트뿐만 아니라 이미지까지 활용하

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

여 사용자의 관심사를 분류하는 방법을 제안하며, 같은 관심사를 사용자 간에 거리 유사도를 측정하여 같은 관심사를 갖고 가장 유사한 관심사를 갖는 사용자를 팔로잉 추천하는 방법에 관해 기술한다.

이어지는 절에서는 본 연구와 관련하여 텍스트와 이미지를 모두 활용하여 관심사를 분류하고 팔로잉 추천을 하기 위한 기반이 되는 기법에 관해 기술한다.

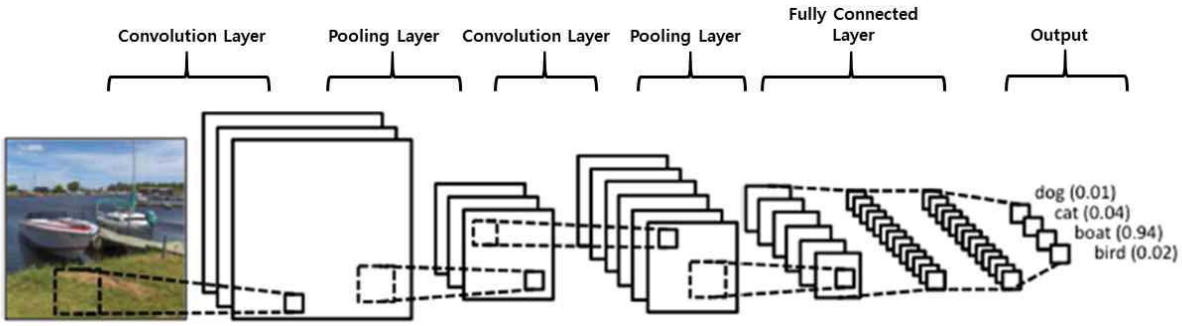
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. Convolutional Neural Network

CNN(Convolutional Neural Network)는 수많은 기계학습 중 하나이며, Neural Network를 기반으로 하는 특징 추출과 인식에 특화된 기법이다. CNN은 1950년대에 고양이의 뇌가 사물을 인식할 때 뇌의 일부분만 활성화되는 모습에서 영감을 얻었다[13, 14]. 1980년대에 들어서 사람이 쓴 글씨를 컴퓨터에 인식시키기 위한 방법으로 CNN의 개념이 도입되기 시작했지만 적용하는데 어려움이 존재했다[15]. 이후에 일반화와 단순화를 거치며 개선되었으나[16, 17] 낮은 컴퓨팅 성능 문제로 여전히 사용이 쉽지 않았다. 최근 컴퓨팅 성능이 개선되고 빅데이터 처리의 중요성이 주목받기 시작하면서 CNN을 비롯한 기계학습이 주목받고 있다[18].

CNN은 실제 뇌가 이미지를 인식하는 원리와 같은 방식으로 작동한다. 뇌가 사물을 인식할 때처럼 이미지를 작은 구역으로 나눠서 각 영역의 특징을 이용해서 학습한 후 입력되는 이미지를 분류해낸다. CNN은 구현하는 방식에 따라 각기 다른 구조와 성능을 보이며, CNN의 기본적인 구성은 [그림 2-3]과 같다.

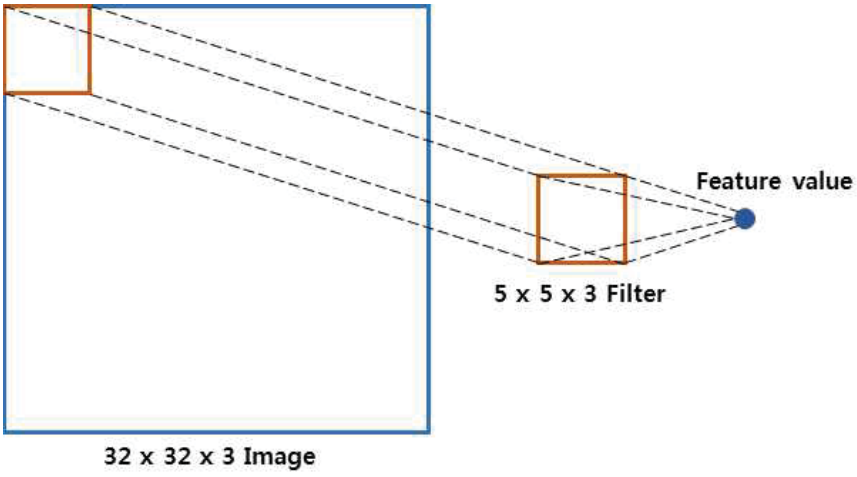


[그림 2-3] CNN 기본 구조

[그림 2-3]에서 알 수 있듯이 CNN은 크게 Convolution layer와 Pooling layer, FC(Fully Connected) layer로 나뉜다. Convolution layer에서 filter를 이용하여 이미지 일부분의 특징값을 추출하고 그 결과를 이용하여 channel로 변환하는 과정을 거친다. 이미지의 시작부터 끝까지 특징값을 추출하는 과정을 거쳐 channel을 생성하는 과정을 반복하여 여러 개의 channel을 생성한다. Convolution layer를 통해 channel을 생성하는 과정은 [그림 2-4]와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 2-4] Convolution layer에서 특징 값 추출

[그림 2-4]는 이미지에 filter를 이용해서 특징값을 추출하는 방법을 나타내고 있다. 입력하는 이미지의 크기에 따라 초기 이미지 크기는 달라질 수 있으며, filter의 크기 또한 설정하고자 하는 값에 따라 다르게 설정할 수 있다. [그림 2-4]에서는 이미지의 크기는 $width \times hight \times depth$로 나타낼 수 있으며, width는 너비, hight는 높이, depth는 색상 값을 뜻한다. [그림 2-4]에서는 색상을 갖는 이미지의 특징값을 추출하기 때문에 3의 값을 가지며, gray scale의 경우 depth의 값은 1을 갖게 된다. 이미지의 크기를 나타내는 방법과 동일하게 filter의 크기를 나타내고 depth의 크기는 이미지의 크기와 동일하게 설정하는 것이 일반적이며, [그림 2-4]에서는 width, hight, depth를 각각 5, 5, 3으로 설정한 것을 알 수 있다.

filter로 설정한 영역으로부터 읽어온 값을 하나의 값으로 변환한다. 식 (1)을 이용하여 특징값을 계산한다.

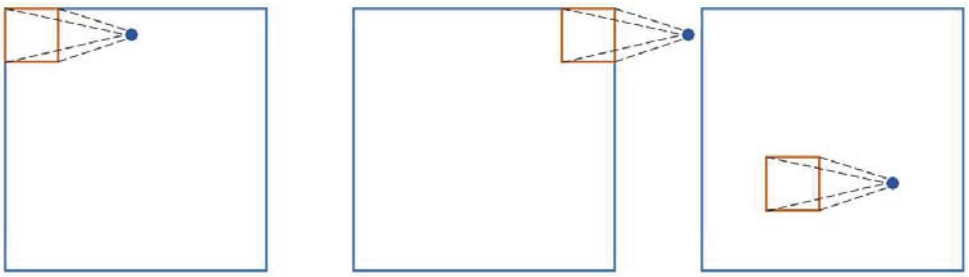
$$H(X) = WX + b \tag{1}$$

식 (1)은 여러 개의 변수를 갖는 linear regression에서 Hypothesis를 계산하기 위한 공식이며, CNN에서 특징값을 계산하기 위해 사용한다. X는 x를 원소로 갖는 집합 또는 행렬을 의미하며, 이미지에 해당하는 값으로 바뀌지 않는다. W는 w를 원소로 갖는 집합 또는 행렬이며, filter에 해당하는 값이다.

filter는 이미지의 끝까지 도달할 때까지 이동하며 각 구역의 특징값을 추출한다. [그림 2-5]는 filter가 이미지의 끝까지 도달하는 과정을 나타낸다.

Collection @ chosun

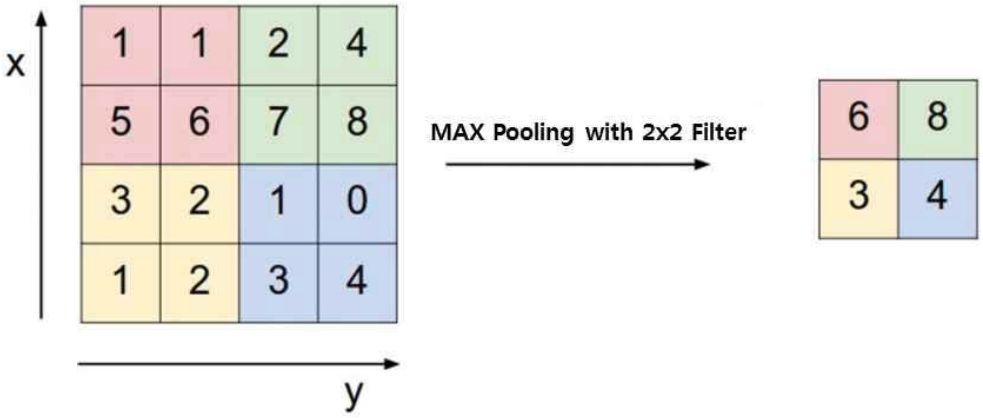
조선대학교
CHOSUN UNIVERSITY



[그림 2-5] Filter의 이동 모습

[그림 2-5]처럼 filter는 왼쪽에서 오른쪽으로 이동하고 오른쪽 끝에 도달했을 경우 아랫줄로 이동하여 다시 왼쪽에서 오른쪽으로 이동하여 이미지의 모든 부분의 특징값을 추출한 후 channel을 생성한다.

Pooling layer는 sampling 또는 resizing이라고 부르기도 하며, Max-Pooling을 이용하여 channel에서 특징점만을 추출한 후 크기를 줄이기 위해 사용한다. Pooling을 이용하여 filter의 크기를 줄이는 방법에는 최소값을 이용하는 방법, 평균을 이용하는 방법, 표준편차를 이용하여 계산하는 방법 등이 있지만, CNN에서는 최대값을 이용하는 Max-Pooling이 가장 많이 이용된다[19]. [그림 2-6]은 Pooling layer에서 Max-Pooling을 적용하는 방법에 대한 과정이다.



[그림 2-6] channel에서 Max-Pooling 이용의 예

[그림 2-6]은 filter의 width와 hight가 2일 경우 Max-Pooling을 적용하는 방법을 나타낸다. filter를 이용하여 4개의 영역으로 구분하고 각 영역에서 가장 큰 값을 선택하여 결과값을 나타낸다.

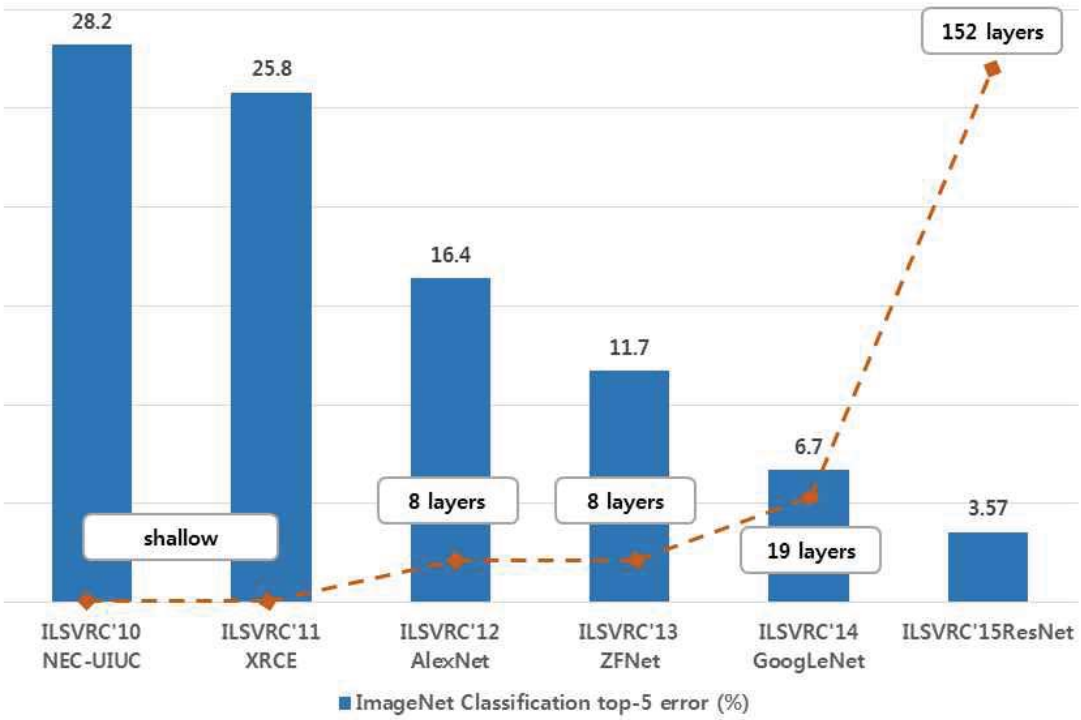
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

앞서 설명한 Convolution layer와 Pooling layer를 반복하여 여러 개의 channel을 생성한 뒤 FC layer에 접어든다. FC는 이름에서 알 수 있듯이 이 전의 모든 layer와 연결되어 있다는 뜻이며, 최종적으로 이미지를 분류하는 역할을 수행한다. 이전 layer에서 추출된 높은 수준의 특징을 이용하여 Softmax를 통해 최종 분류 결과를 결정하게 되며, [그림 2-3]에서는 FC layer에서 dog, cat, boat, bird 중에서 하나로 이미지를 분류해냄을 보인다..

1. 이미지 분류

CNN을 구현하여 이미지를 학습하고 올바르게 분류하기 위한 다양한 연구들이 진행되고 있으며, ImageNet은 수백 개의 객체 카테고리와 수백만 개의 이미지를 대상으로 이미지 분류와 기준을 확립하기 위해 ILSVRC(Imagenet Scale Visual Recognition Challenge)를 주최하고 있다[20]. ILSVRC는 2010년부터 시작되었고 기계 학습을 이용한 대규모 이미지 분류를 위해 매년 다양한 팀들이 참가하고 있으며, [그림 2-7]은 ILSVRC의 역대 우승팀과 top-5 error rate, layer의 개수를 보여준다[21,22,23,24].



[그림 2-7] ILSVRC의 우승팀, top-5 error rate, layer의 개수

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2010년과 2011년은 CPU를 이용한 이미지 분류를 목적으로 대회가 개최되었기 때문에 얕은 layer를 가지며 높은 error rate를 보인다. 2012년부터 GPU를 사용하기 시작했으며 layer 또한 증가하기 시작하고 점차 낮은 error rate를 보인다. [그림 2-7]을 통해 layer가 많아질수록 낮은 error rate를 보이며, 좋은 성능의 CNN을 생성할 수 있다는 것을 알 수 있다. AlexNet은 2개의 GPU를 기반으로 한 병렬 구조로 구성되어 있으며 총 5개의 Convolution layer와 3개의 FC layer로 구성되어 있다. 처음으로 GPU를 사용하여 이미지를 분류하였으며, 압도적인 성능으로 1위를 차지함으로써 크게 주목을 받았다[21]. 이후 GPU를 이용한 다양한 CNN 모델들이 연구되었으며 더 좋은 결과를 낼 수 있게 되었다. ILSVRC 2013에서 우승한 ZFNet은 11.7%의 error rate를 보이며 우승을 한다. ZFNet은 새로운 CNN 모델을 구축한 것이 아니라 AlexNet에 Visualizing 기법을 적용함으로써 CNN을 더 잘 이해할 수 있도록 하고 기존의 CNN 모델을 개선하는 것에 중점을 두었다. 그로 인해 기계학습 분야의 연구자들에게 더 좋은 CNN 구조를 설계할 수 있도록 큰 영향을 미쳤다[22]. ILSVRC 2014에서 우승한 CNN 모델은 GoogLeNET이며, 이때부터 CNN 구조에 큰 변화가 나타나기 시작한다. 기존의 CNN 모델에 비해 구조가 단순해지고 CNN의 전형적인 형태를 보이기 시작하며, layer를 증가시킴으로써 더 깊이 있는 Network를 구축하기 시작한다. 깊이 있는 CNN 모델을 구축하기 위해서는 불가피하게 layer를 증가시킬 수밖에 없으며, 그에 따른 연산량이 증가할 수밖에 없지만, width와 hight가 1의 값을 갖는 convolution을 적절히 사용하여 연산량이 증가하는 문제를 해결했다[23]. ILSVRC 2015에서 우승한 CNN 모델은 ResNET이다. ResNET은 Residual learning, Shortcut Connection, Identity 등 새로운 개념을 도입하였으며, layer 수를 152개까지 증가시키는 ultra-deep의 구조를 갖는다. 기존의 CNN은 20개 내외의 layer로 구성되어 있으며, layer의 수가 그 이상으로 증가하는 경우 성능이 개선되지 않은 한계점이 있었다. RestNet은 layer의 수가 증가하여 Network가 깊어질 경우 학습 속도가 크게 저하되는 문제와 학습 효과가 사라지는 문제를 개선하여 3.57의 낮은 error rate를 기록했다[24].

본 논문에서는 적당한 깊이를 갖는 20개 내외의 layer로도 뛰어난 성능을 보이며, 현실적으로 구현하여 사용 가능한 GoogLeNet을 이용하여 SNS에 적합한 이미지를 학습시키고 관심사 카테고리를 기반으로 이미지를 분류한다.

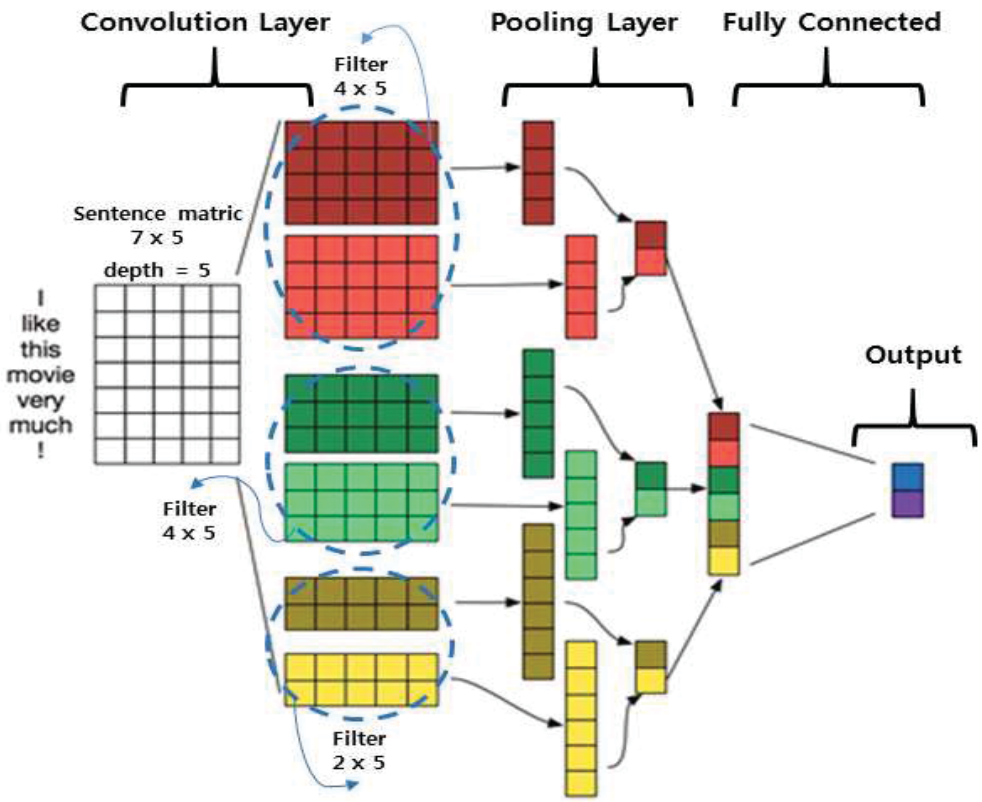
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 텍스트 분류

앞서 설명한 것처럼 CNN은 필기체를 컴퓨터가 정상적으로 인식하도록 하고자 하는 연구에서 시작되었기 때문에 이미지 분류를 위한 연구를 중점으로 진행되었으나 최근에는 CNN을 이용하여 텍스트를 분류하는 것이 좋은 성능을 보인다는 것이 증명됨으로써 CNN을 이용한 텍스트 분류에 관한 연구 또한 활발히 진행되고 있다[25,26,27,28].

텍스트를 분류하기 위해 CNN을 사용할 경우 이미지를 분류할 때와 같은 구성을 갖지만 입력 데이터가 텍스트이기 때문에 filter와 channel이 약간 다른 모습을 보이며, filter의 크기를 하나로 설정하는 것이 아니라 여러 개로 설정한다. 일반적으로 2, 3, 4의 크기를 갖는 필터를 각각 2개씩 갖는다. [그림 2-8]은 텍스트를 분류하기 위한 CNN의 구성이다.



[그림 2-8] 텍스트 분류를 위한 CNN 구성

[그림 2-8]에서 알 수 있듯이 기본적인 Convolution layer와 Pooling layer, FC

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

layer가 존재하는 것을 알 수 있다. 입력 데이터가 다르기 때문에 문장을 단어 단위로 토큰화시킨 후 행렬의 각 행에 문장의 각 단어를 벡터 형식으로 입력한다는 점이 이미지를 분류할 때와 다르다. 위 그림에서는 “I like this movie very much !”를 단어 단위로 토큰화시켰으며, depth를 5로 설정했음을 알 수 있다. 2, 3, 4의 크기를 갖는 필터를 각각 2개씩 생성하였으며, 특징값을 갖는 channel을 생성한다. 1의 크기를 갖는 filter를 생성하여 Max-Pooling을 거쳐 가장 큰 값을 추출하여 channel의 크기를 줄인다. FC layer에서 softmax를 이용하여 텍스트를 분류하며, 위 그림에서는 2개의 결과물로 분류한다.

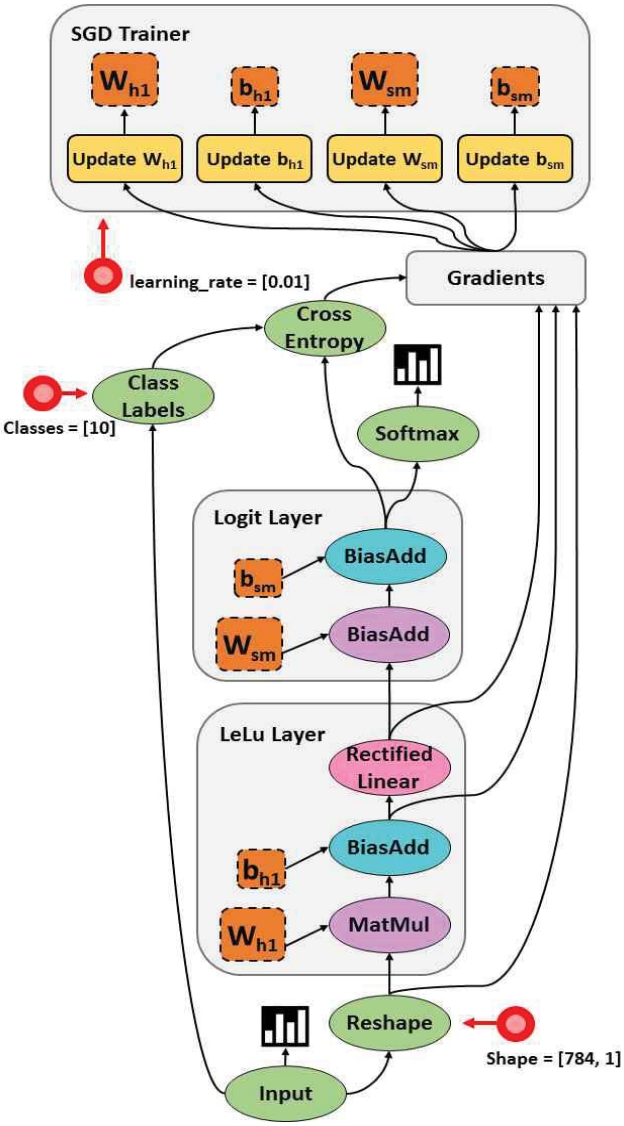
본 논문에서는 CNN을 이용하여 텍스트를 분류하기 위해 [25]을 참고하여 CNN을 구현한다[29,30]. [25]은 CNN을 이용한 텍스트 분류의 초석이 되는 연구였으며, 매우 간단하고 강력한 성능을 자랑하는 Network를 구성하고 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. TensorFlow

TensorFlow는 기계학습과 딥러닝을 위해 Google에서 만든 오픈 소스 라이브러리이며, 데이터 플로우 그래프 방식을 사용하여 수치 연산이 가능하고 Python과 C++을 통해 프로그램 구현이 가능하다. 유연한 시스템으로 구성되어 있기 때문에 약간의 코드 수정으로도 다양한 기기에서 CPU 또는 GPU를 사용하여 연산이 가능하다. TensorFlow에서 데이터는 tensor로 표현하고 tensor는 다차원의 동적 배열이며, tensor가 플로우 그래프에서 흐르는 것처럼 작동한다. [그림 2-9]는 텐서플로우 그래프의 예이다[31,32,33].



[그림 2-9] 텐서 플로우 그래프의 예

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

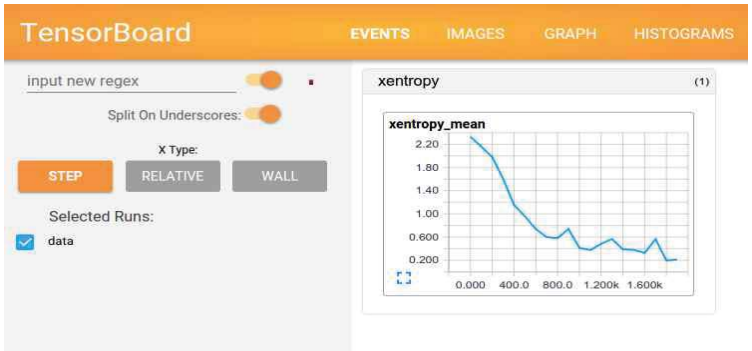
[그림 2-9]에서 알 수 있듯이 그래프가 흐르는 것처럼 구성되어 있으며, 노드와 엣지를 사용한 방향 그래프로 표현되어 있다. 노드는 수학적 계산과 데이터 입력, 출력, 데이터 읽기, 저장 등의 작업을 수행하며, 엣지는 노드 사이의 데이터 입출력 관계를 나타낸다.

TensorFlow를 이용하여 프로그램을 구현하기 위해서는 TensorFlow에서 사용하는 용어를 사용해야 하며, 이를 통해 미분, 적분 등의 어려운 연산이나 deep Neural Network 등의 복잡한 구조를 갖는 프로그램을 구현할 수 있다. [표 2-1]은 TensorFlow에서 사용하는 용어를 나타낸다.

[표 2-1] TensorFlow 용어

Term	Contents
Operation(Op)	그래프 상의 노드를 나타내며. 계산을 수행
Tensor	타입을 가진 다차원 배열으로 내부적인 데이터를 전달
Session	그래프를 실행하기 위한 객체
Variables	그래프 실행 시 파라미터를 저장하고 갱신
Run	그래프에서 연산을 실행
Session	그래프에서 연산을 실행하기 위한 함수 제공
Device	연산을 실행할 기기(GPU 또는 CPU) 설정 및 실행

TensorFlow에서는 deep Neural Network 훈련 같은 복잡한 연산을 할 경우 도출되는 결과와 적합한 결과를 거쳤는지 이해하기 쉽게 도우며, 구현한 프로그램을 최적화하기 위해 TensorBoard라는 시각화 도구를 제공한다. TensorBoard는 코드 실행에 대한 정량적인 측정 결과를 표시해주며, [그림 2-10]과 같다.



[그림 2-10] TensorBoard의 예

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

D. DMOZ

웹은 놀랄만한 속도로 계속된 성장을 하고 있으며, 자동화된 검색 엔진은 최적화된 유용한 검색 결과를 제공하는 부분에서 한계가 있다. 폭발적으로 증가하는 웹사이트를 질적인 부분과 양적인 부분에서 효과적으로 정리할 수단이 부족한 실정이다. 이를 개선하기 위하여 디렉토리를 기반으로 웹사이트를 정리하기 위한 프로젝트가 진행되고 있다. DMOZ는 월드 와이드 웹 링크의 다국적 ODP(Open Directory Project)로 진행되는 프로젝트이며, 전 세계의 자발적인 편집자 모임으로 구성된 다양한 사용자의 참여로 구축되는 디렉토리 형식의 카테고리이다. ODP는 사용자가 직접 분류한 디렉토리 기반 웹 콘텐츠 데이터베이스로 Google, Lycos, DirectHit 등의 검색 엔진과 포털에 웹 사이트의 정보를 제공하고 있다[34].

DMOZ는 Arts, Business, Computers, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports, Kids & Teens, DMOZ around the World의 총 16개의 카테고리로 구성되어 있으며, [그림 2-10]과 같다.

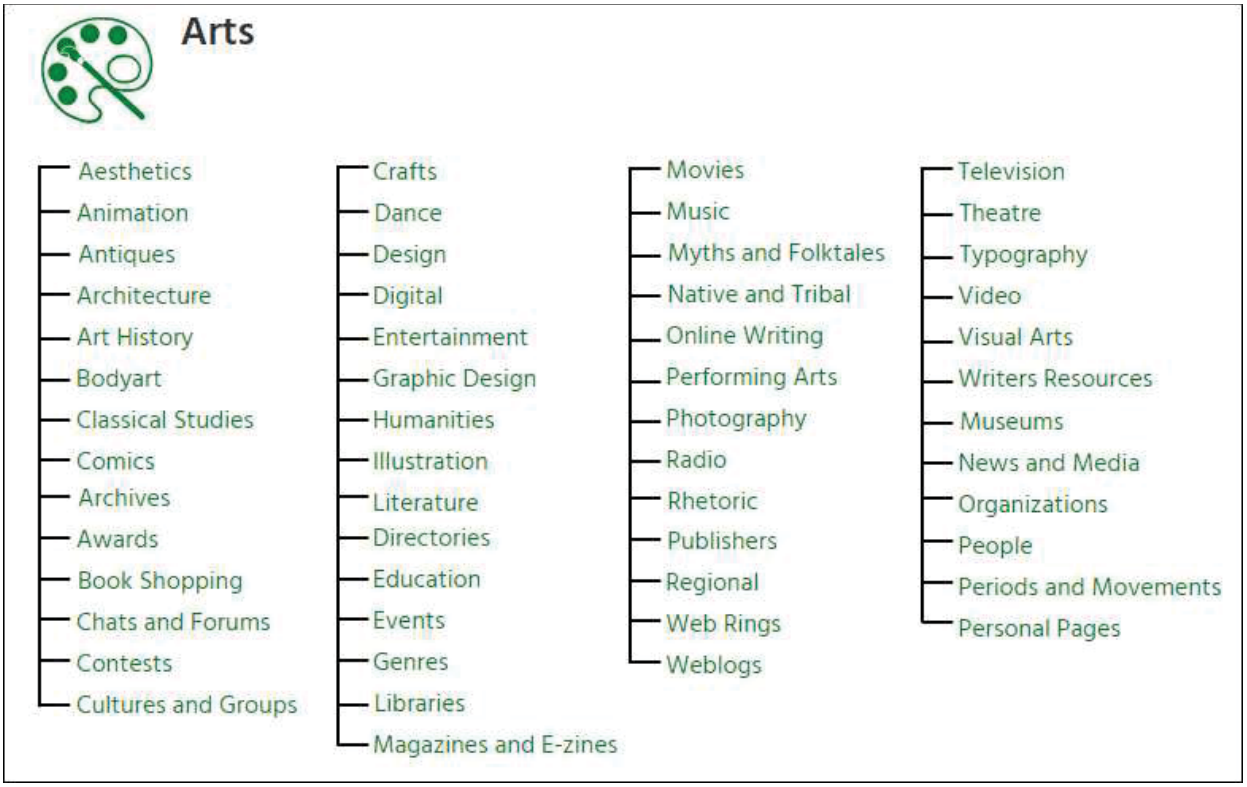
[그림 2-11] DMOZ 카테고리

DMOZ는 디렉토리 형식으로 구성되어 있으며, 최상위 디렉토리부터 최하위 디렉토리까지 실제 디렉토리처럼 상위와 하위 구조의 트리 형태를 가지고 있으며, 관련

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

있는 카테고리 또한 표시해준다. [그림 2-11]은 DMOZ에서 제공하는 디렉토리 형태를 보여준다.



[그림 2-12] DMOZ 카테고리의 형태

웹을 사용하는 사람들이 직접 참여하여 만들어지는 카테고리인 만큼 사람의 생각이 반영된 객관적인 카테고리의 기준이라고 판단하여 본 논문에서는 DMOZ에서 제공하는 카테고리를 사용자 관심사 카테고리로 재정의하고 분류하는 기준으로 사용한다.

Collection @ chosun

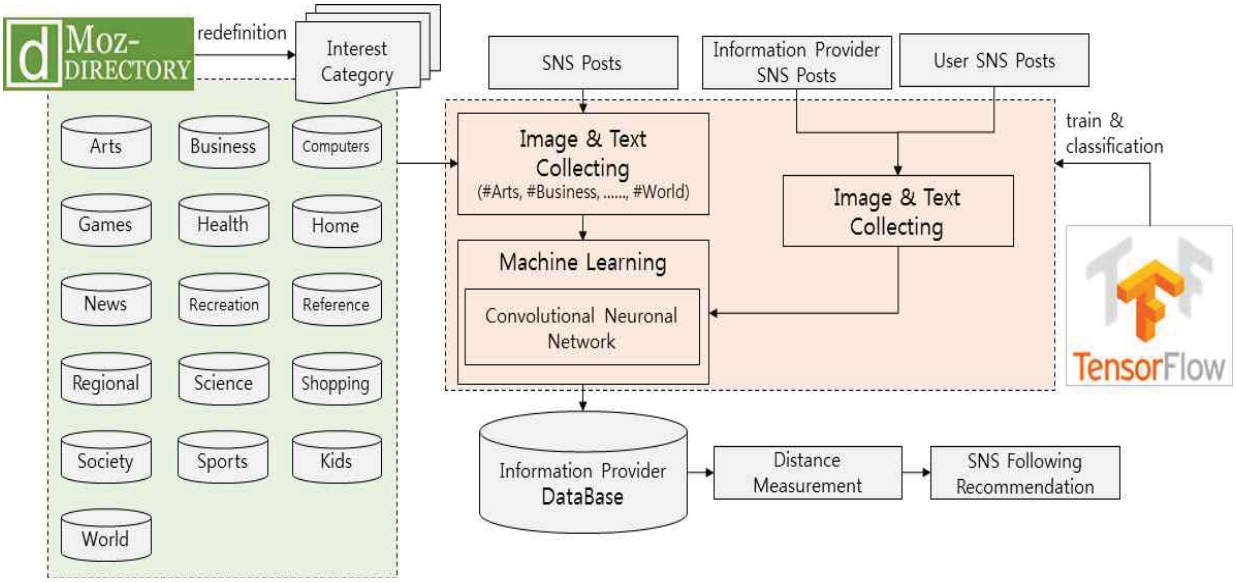
조선대학교
CHOSUN UNIVERSITY

Ⅲ. 관심사 카테고리 분류를 이용한 팔로잉 추천 방법

본 논문에서는 SNS 게시글의 텍스트와 이미지를 함께 이용하여 CNN을 통해 DMOZ를 기준으로 하는 관심사 카테고리를 분류하고 비슷한 관심사를 갖는 사용자 간에 팔로잉을 추천하는 방법에 대해 제안한다.

A. 시스템 구성도

[그림 3-1]은 본 논문에서 제안하는 관심사 카테고리 분류를 이용한 팔로잉 추천 방법의 전체 시스템 구성도이다.



[그림 3-1] 전체 시스템 구성도

본 논문에서 제안하는 관심사 카테고리 분류를 이용한 팔로잉 추천 방법은 크게 CNN에 학습할 SNS 게시글의 이미지와 텍스트를 학습하는 과정과 정보제공자 데이터베이스를 구축하는 과정, 사용자의 관심사 카테고리를 분류하여 팔로잉 추천하는 과정으로 구성되어 있다.

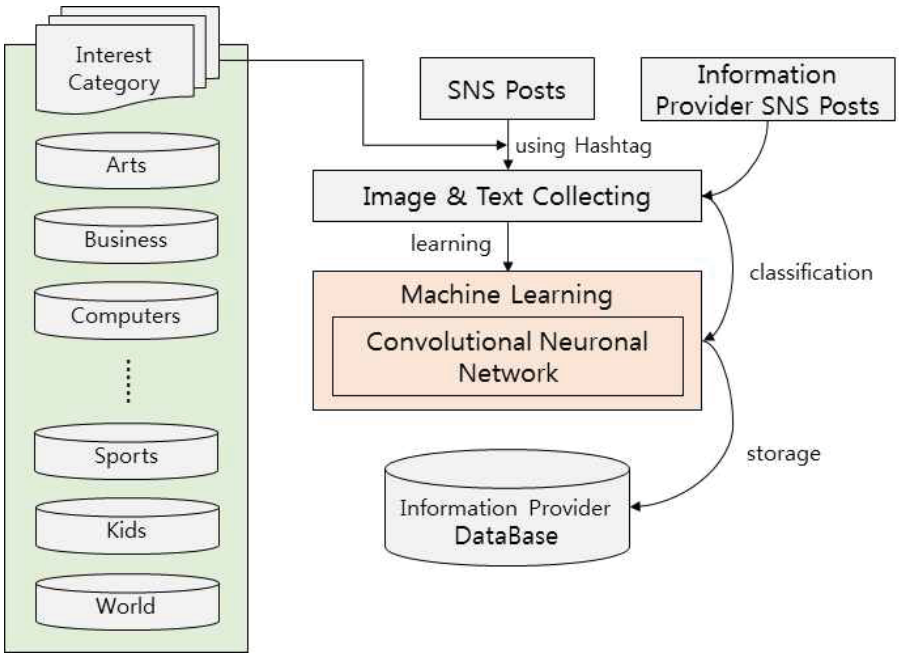
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 정보제공자 데이터베이스 구축

SNS 사용자는 서로 관계를 형성하고 있으며, 정보나 서로의 의견을 공유하기 위해 다양한 게시글을 활용하고 있고 성향에 따라 다양한 정보를 게시하는 것에 몰두하는 사용자와 게시글을 관찰하는데 몰두하는 사용자로 나뉜다. 본 논문에서는 정보 공유에 필요한 다양한 정보를 게시하며, 정보를 전달하는 역할을 하는 사용자를 정보제공자로 정의한다. 따라서 100개 이상의 글을 게시한 사용자를 정보제공자로 판단한다.

본 절에서는 사용자에게 팔로잉 추천을 해주기 위한 대상이 되는 정보제공자의 정보를 가지고 있는 정보제공자 데이터베이스를 구축하는 방법에 관해 기술한다. [그림 3-2]는 정보제공자 데이터베이스 구축의 흐름을 나타내는 구성도이다.



[그림 3-2] 정보제공자 데이터베이스 구성도

정보제공자 데이터베이스를 구축하기 위해 관심사 카테고리를 정의하고 CNN을 이용하여 SNS 게시글의 이미지와 텍스트를 학습한다. 학습을 통해 구축된 모델을 이용하여 정보제공자 SNS 게시글의 이미지와 텍스트를 입력하고 관심사 카테고리를 분류하여 정보제공자 데이터베이스를 구축한다. 정보제공자에게 ID와 이미지와 텍스트를 분류할 때 도출된 값을 부여하여 정보제공자 데이터베이스에 저장한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

1. 관심사 카테고리 정의

DMOZ는 Arts, Business, Computers, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports, Kids & Teens, DMOZ around the World의 총 16개의 카테고리로 구성되어 있으며, 최상위 카테고리를 나타낸다. Kids & Teens의 경우 결합된 단어 형태를 띠고 있기 때문에 간편하게 표기하기 위해 Kids로 나타내며, DMOZ around the World의 경우 World와 같은 뜻으로 해석할 수 있기 때문에 World로 재정의 하며, 관심사 카테고리로 정의한다. [표 3-1]은 DMOZ의 최상위 카테고리를 재정의한 관심사 카테고리를 나타낸다.

[표 3-1] 관심사 카테고리

No	Category
1	Arts
2	Business
3	Computers
4	Games
5	Health
6	Home
7	News
8	Recreation
9	Reference
10	Regional
11	Science
12	Shopping
13	Society
14	Sports
15	Kids
16	World

[표 3-1]의 관심사 카테고리를 기준으로 이미지와 텍스트를 분류하여 SNS 사용자의 관심사를 분류한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. CNN을 이용한 관심사 카테고리 분류

본 절에서는 정보제공자 데이터베이스를 구축하기 위해 SNS 게시글에서 이미지와 텍스트를 입력데이터로 하여 관심사 카테고리를 분류하는 방법에 관해 기술하며, 크게 이미지와 텍스트에서 각각 관심사 카테고리를 분류하는 단계와 분류된 결과를 통해 정보제공자 데이터베이스를 구축하는 단계로 나뉜다.

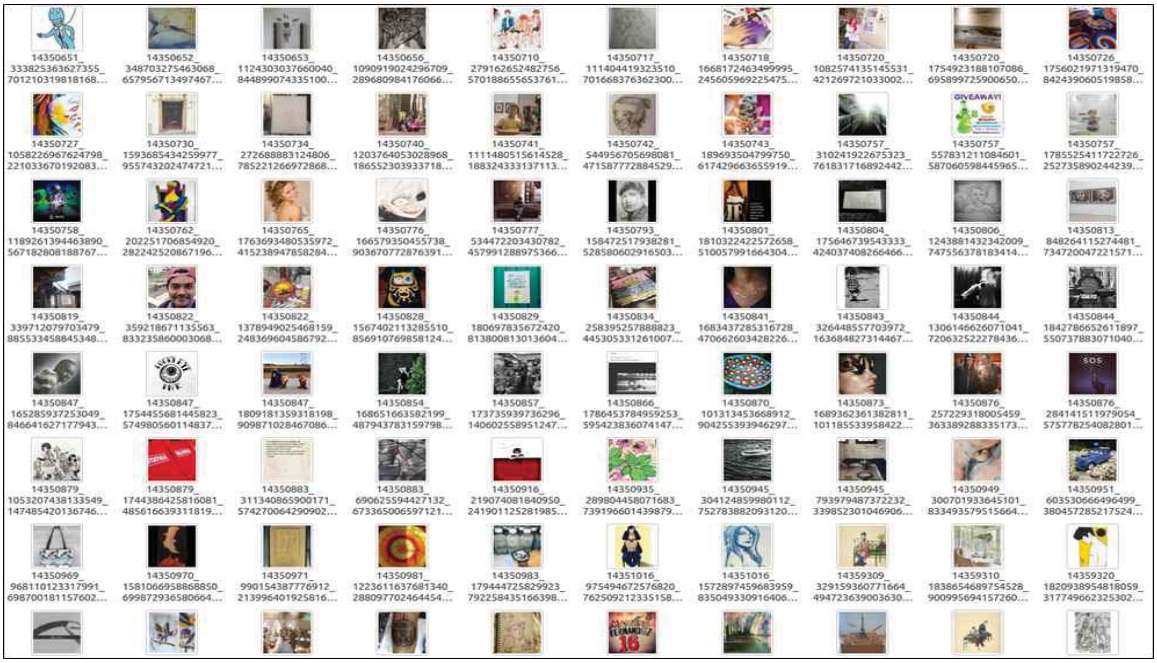
a. 이미지에서 관심사 카테고리 분류

이미지에서 관심사 카테고리를 분류하기 위해 학습 이미지 수집 단계, CNN 구축 및 학습 모델 생성 단계, 정보제공자 이미지 수집 단계, CNN 학습 모델을 이용한 분류 단계를 거친다.

학습 이미지를 수집하기 위해서 SNS 중 인스타그램을 이용한다. 인스타그램은 사진 중심으로 글을 게시하는 SNS이며, 사진과 함께 간단한 문구와 해시태그를 이용하여 게시글을 작성하고 팔로잉을 통해 다른 사용자들과 함께 정보를 공유한다[35, 36]. 따라서 다양하고 방대한 양의 이미지 정보를 제공하는 SNS인 인스타그램에서 학습하고자 하는 이미지를 수집하며, 관심사 카테고리를 기준으로 해시태그를 검색하여 카테고리별로 이미지를 1000개씩 수집한다. 인스타그램 게시글의 내용을 습득하기 위해서 기존에는 개발자를 위한 API(Application Programming Interface)를 사용할 수 있었다. 하지만 2015년 11월 17일에 발표한 인스타그램 API 정책 변경으로 인해 게시글의 사진을 불러와서 보여주거나 특정 해시 태그의 피드를 가져와서 보여주는 경우 등 게시글 내의 내용에 접근할 수 없게 되었으며, 인스타그램 API를 사용하여 개발한 프로그램 중에서 Insta 또는 Gram 등의 명칭을 사용하며 인스타그램으로 오해를 살 수 있는 프로그램에 대해서도 API 제공을 중지하기로 했다. 다만 인스타그램 게시글에 접근하고자 하는 프로그램을 제작하고자 할 경우 프로그램 심사 신청을 하고 통과한 후 사용이 가능하다[37]. 따라서 이미지를 수집하기 위해 크롬 확장프로그램인 Fatkun Batch Download Image를 사용한다. Fatkun은 크롬 웹 스토어에서 제공하며, 크롬에 표시되는 모든 이미지를 간단하고 편리하게 다운받을 수 있다[38]. [그림 3-3]은 #Arts를 해시태그로 갖는 게시글에서 이미지를 수집한 결과이다.

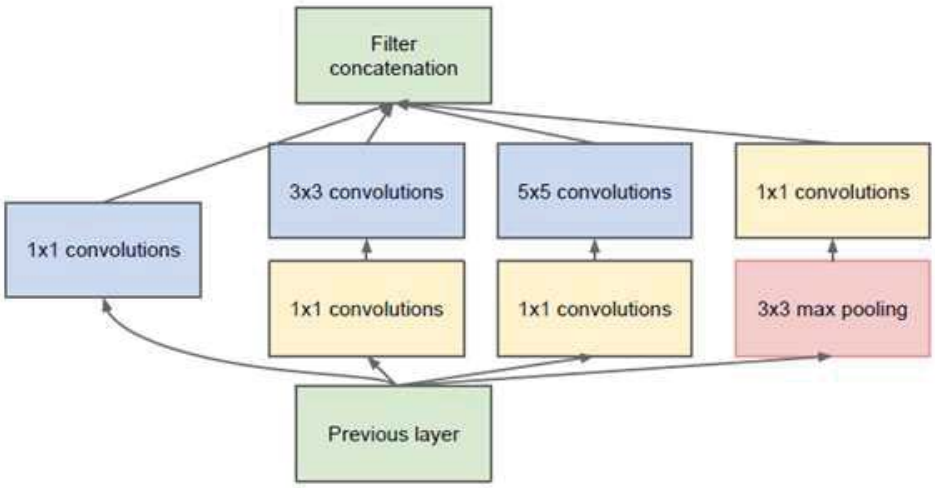
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 3-3] 학습을 위해 수집한 이미지의 예

인스타그램에서 수집한 이미지를 학습하기 위해 TensorFlow를 이용하여 CNN 모델을 구축한다. TensorFlow를 이용하여 구현한 프로그램은 수정이 간편하고 TensorBoard를 통해 프로그램의 흐름과 연산 과정, 결과를 손쉽게 활용할 수 있다. TensorFlow에서 제공하는 Inception-v3 모델을 수정하여 CNN을 구축하였다. Inception-v3 모델은 ILSVRC 2014에서 구글이 우승할 당시의 CNN 모델인 GoogLeNet의 한 종류이다.



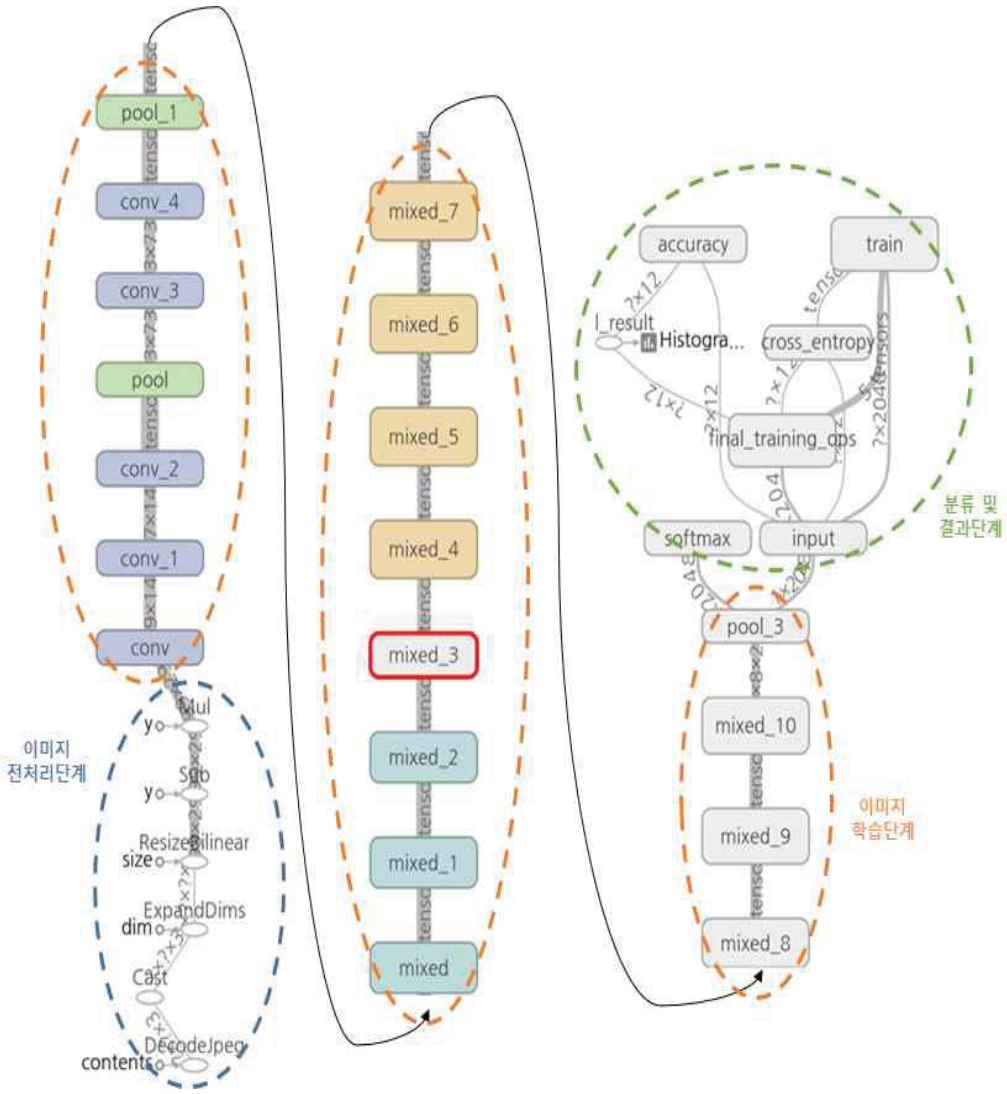
[그림 3-4] Inception-v3의 기본 구조

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-4]의 Inception-v3의 기본 구조에서 알 수 있듯이 다양한 크기의 filter를 갖는 convolution을 여러 번 수행하는 구조로 되어 있으며, 같은 layer에 서로 다른 크기를 갖는 filter를 적용하여 망이 깊어짐에 따라 발생하는 연산량 문제를 해결했다[39, 40].

본 논문에서는 이미지를 관심사 카테고리로 분류하기 위해 TensorFlow를 이용하여 Inception-v3를 수정하여 프로그램을 구현했다. Inception-v3의 기본 구조를 따르지만, SNS 환경에 적합한 이미지 데이터를 학습시키기 위해 인스타그램에서 수집한 이미지를 이용하여 학습시킴으로써 학습 모델을 생성하며, 일종의 라벨링이라고 할 수 있는 최종적으로 분류되는 결과를 관심사 카테고리로 도출될 수 있도록 했다. [그림 3-5]는 본 논문에서 사용하는 CNN 모델의 플로우 차트이다.



[그림 3-5] 이미지 분류를 위한 CNN 모델의 플로우 차트

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-5]에서 conv 노드는 Convolution layer를 나타내고 pool 노드는 Pooling layer를 나타낸다. mixed 노드는 앞서 [그림 3-4]에서 설명한 Inception-v3의 기본 구조를 가지고 있다. 플로우 차트의 흐름을 살펴보면 conv 노드 도달 이전에는 이미지의 포맷를 jpeg의 형태로 일치시키고 이미지의 크기를 하나로 일치 시키는 전처리가 수행된다. 그 이후 5개의 conv 노드와 3개의 pool 노드, 11개의 mixed 노드를 통해 학습하며, softmax 노드로 이동하여 16개의 관심사 카테고리 중 하나로 분류되거나 input 노드로 이동하여 학습에서 발생한 값을 가공하여 다양한 연산 결과를 도출해낸다. cross_entropy 노드에서는 손실률을 계산하고 accuracy 노드에서는 정확률을 계산한다. [그림 3-6]은 [그림 3-3]에 해당하는 인스타그램 이미지를 학습 데이터로 이용하여 본 논문에서 구현한 CNN으로 학습한 결과이다.

```
2016-10-26 20:12:53.355588: Step 3930: Validation accuracy = 26.0%
2016-10-26 20:12:53.974656: Step 3940: Train accuracy = 40.0%
2016-10-26 20:12:53.974720: Step 3940: Cross entropy = 1.788089
2016-10-26 20:12:54.035492: Step 3940: Validation accuracy = 32.0%
2016-10-26 20:12:54.648391: Step 3950: Train accuracy = 42.0%
2016-10-26 20:12:54.648457: Step 3950: Cross entropy = 1.661950
2016-10-26 20:12:54.708017: Step 3950: Validation accuracy = 30.0%
2016-10-26 20:12:55.319804: Step 3960: Train accuracy = 43.0%
2016-10-26 20:12:55.319868: Step 3960: Cross entropy = 1.840439
2016-10-26 20:12:55.377700: Step 3960: Validation accuracy = 37.0%
2016-10-26 20:12:55.993815: Step 3970: Train accuracy = 40.0%
2016-10-26 20:12:55.993882: Step 3970: Cross entropy = 1.815057
2016-10-26 20:12:56.050870: Step 3970: Validation accuracy = 39.0%
2016-10-26 20:12:56.667216: Step 3980: Train accuracy = 49.0%
2016-10-26 20:12:56.667280: Step 3980: Cross entropy = 1.668399
2016-10-26 20:12:56.725288: Step 3980: Validation accuracy = 27.0%
2016-10-26 20:12:57.345691: Step 3990: Train accuracy = 40.0%
2016-10-26 20:12:57.345755: Step 3990: Cross entropy = 1.783113
2016-10-26 20:12:57.405997: Step 3990: Validation accuracy = 41.0%
2016-10-26 20:12:57.961544: Step 3999: Train accuracy = 42.0%
2016-10-26 20:12:57.961610: Step 3999: Cross entropy = 1.888347
2016-10-26 20:12:58.020140: Step 3999: Validation accuracy = 36.0%
```

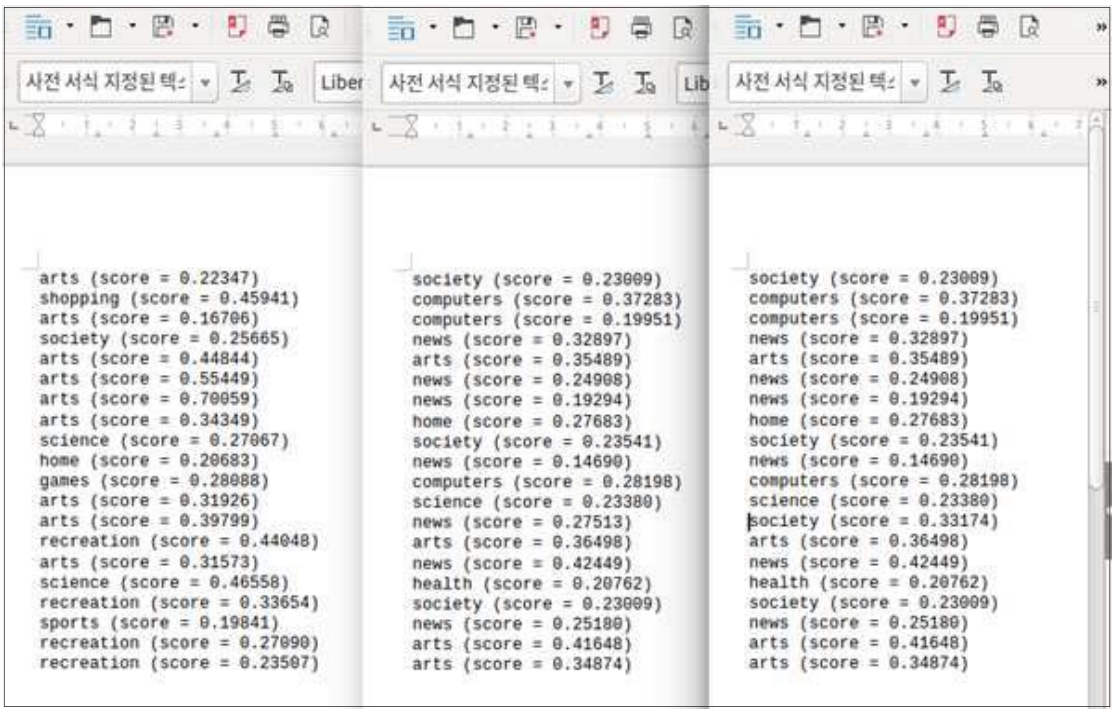
[그림 3-6] 이미지 분류를 위한 CNN 모델의 학습 결과

이미지 분류를 위한 CNN 모델을 학습하는 과정을 나타내고 있다. 총 3,999단계를 거쳐 학습 모델을 구축하였으며, 정확률이 얼마나 정확한지 나타내는 Validation accuracy와 학습이 얼마나 정확하게 되었는지 나타내는 Train accuracy, 학습 중에 발생하는 이미지의 손실률이 얼마나 되는지 나타내는 Cross entropy의 값이 함께

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

출력되는 것을 알 수 있다. 학습된 모델에 정보제공자에 해당하는 사용자의 이미지를 입력하여 관심사 카테고리를 기준으로 관심사를 분류했으며, 카테고리마다 10명의 정보제공자를 선정했고 각 20개의 이미지를 수집하여 총 3,200개의 이미지를 분류했다.



[그림 3-7] 학습 모델을 이용하여 분류한 이미지 결과의 예

[그림 3-7]은 세 명의 정보제공자의 SNS 게시글에서 각 20개의 이미지를 수집하여 학습 모델을 통해 관심사 카테고리를 분류한 결과이며, 결과 값으로 분류된 카테고리 명과 입력된 이미지와 분류한 이미지가 얼마나 유사한지 나타내주는 score 값으로 구성되어 있다. 분류하기 위해 입력한 이미지에서 정확하게 분류된 이미지의 socre값의 평균을 측정하여 가장 유사한 정보를 가진 정보제공자를 사용자에게 추천할 때 사용한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

b. 텍스트에서 관심사 카테고리 분류

텍스트에서 관심사 카테고리를 분류하기 위해 학습 테스트 수집 단계, CNN 구축 및 학습 모델 생성 단계, 정보제공자 텍스트 수집 단계, CNN 학습 모델을 이용한 분류 단계를 거친다.

학습 텍스트를 수집하기 위해서 학습 데이터로 적합한 적합한 SNS인 페이스북을 이용한다. 페이스북은 텍스트 중심의 SNS이지만 게시글의 용도에 맞게 이미지도 혼용해서 사용 가능한 SNS이다. 인스타그램과 마찬가지로 텍스트와 사진을 게시할 때 해시태그를 이용하여 작성 가능하며, 팔로잉을 통해 다른 사용자들과 함께 정보를 공유한다[5,7]. 관심사 카테고리를 키워드로 게시글을 검색하여 카테고리별로 텍스트를 1,000문장씩 수집한다. 페이스북 게시글의 내용을 습득하기 위해서 자바 기반의 Facebook API Library인 Facebook4j를 이용한다. Facebook4j를 이용하여 게시글의 내용과 이미지뿐만 아니라 친구 목록, 댓글, 사용자 정보 등에 관련된 내용도 수집할 수 있다[41]. 사용자 게시글을 수집한 후 텍스트를 분류하기 위한 CNN 모델에 사용하기 적합한 형태로 전처리과정을 수행한다. 게시글을 문장단위로 분할하고 키워드에 해당하는 단어를 문장의 끝에 태깅한다. [그림 3-8]은 페이스북에서 수집한 텍스트를 학습에 적합한 형태로 변형한 결과의 예이다.

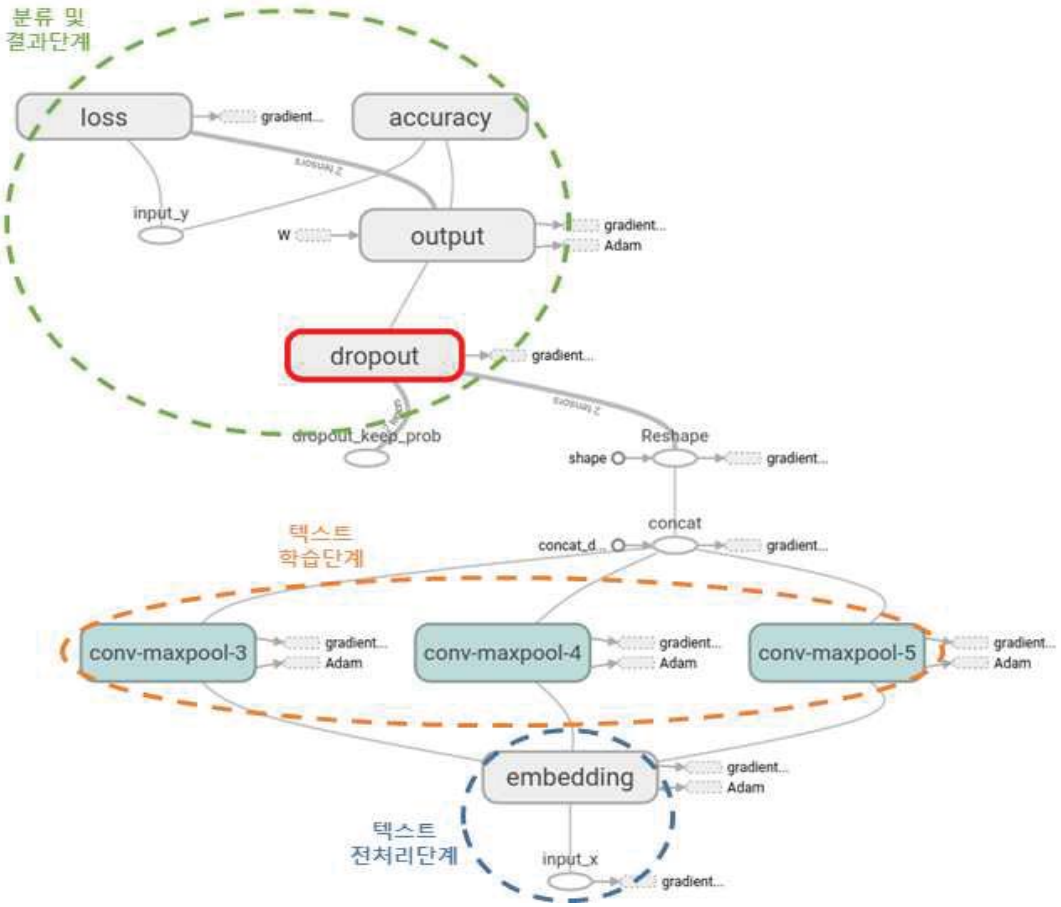
```
PlanSponsor article on retirement income strategies. See the Stanford & SOA report      Society
http://ow.ly/9HUK305jlCL http://ow.ly/nIp2305jmAg       Society
Kai Huang, the first person from China to be granted the SOA's General Insurance track FSA, shares his
How to boost your communication skills in a very short time? Join us in Hartford, Connecticut on Nov. 1
session, as well as one on one coaching opportunities brought to you by Own The Room !  Society
"Own The Room has helped countless speakers play at a higher level," said Sheryl Sandberg, COO of Faceb
Register for this unique opportunity!   Society
Here's how to stop writing emails that bury what you're asking for under a bunch of inessential informa
Retirement shocks focus of Forbes article, references SOA retirement risk survey findings http://ow.ly/
CBS Moneywatch on research from the SOA and the Stanford Center on Longevity on retirement income http:
U.S. News & World Report article highlights new solutions developed by a young student. A 12 year old g
prescription drugs. Read about her insights in address this problem.  http://ow.ly/RP2o305bQjy  Society
This week, the SOA University Outreach team visited Texas A&M University and The University of Texas at
answered questions from the future actuaries.   Society
Subscribe to our Facebook event updates and join us for a university visit near you!    Society
Walter Updegrave on getting the most out of retirement planning. Read about the Actuaries Longevity Ill
http://ow.ly/7dzP3059XQ5 www.longevityillustrator.org   Society
According to Forbes , actuary is one of the most promising jobs for millennials.        Society
Becker's Hospital Review article covers risk adjustments in the ACA individual market. Actuary Rebecca
http://ow.ly/QwKT3055eh4 http://ow.ly/xaq93055evn       Society
How do data science salaries compare? Find out on SOA President Craig Reynolds' SOA LinkedIn page  http
How can you stand out from other qualified candidates in today's entry level job market?        Society
Check out the free recording of the recent SOA Candidate Connect webcast: How Do I Get a Job?   Society
Congratulations to Kai Huang, the first person in Asia to obtain the FSA in General Insurance track! Re
fields. Society
Actuary Steve Tutewohl on applying actuarial skills in health care predictive analytics http://ow.ly/dV
http://ow.ly/C7443040kuH        Society
The Inside Track The Power of Predictive Analytics      Society
```

[그림 3-8] 학습을 위해 수집 후 전처리를 거친 텍스트의 예

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

이미지를 학습한 CNN 모델을 구축했던 것처럼 TensorFlow를 이용하여 텍스트를 학습하기 위한 모델을 구축한다. 텍스트를 분류하기 위한 기본적인 CNN 모델과 동일하게 Convolution layer와 Pooling layer로 이루어져 있다. 이미지 분류에 사용되는 CNN 모델과 가장 큰 차이점은 전처리 과정으로 embedding layer가 필요하다는 점이다. embedding layer는 텍스트를 학습하기 전에 학습하기 적합한 말뭉치 단위로 텍스트를 분할하는 단계로 이미지 CNN 모델의 Convolution layer에서 filter를 정하는 것과 비슷한 과정으로 문장을 단어 단위의 행렬로 변환하며, 그 이후의 과정은 이미지 CNN 모델과 동일하게 Convolution layer와 Pooling layer를 거친 후 FC layer에서 softmax를 이용하여 관심사 카테고리를 분류한다.



[그림 3-9] 텍스트 분류를 위한 CNN 모델의 플로우 차트

[그림 3-9]는 텍스트 분류를 위한 CNN 모델의 플로우 차트이다. 크게 embedding 노드와 conv-maxpool 노드, dropout 노드로 구성되어 있다. 앞서 설명한 대로 embedding 노드에서 입력한 문장 단위의 텍스트를 단어 단위의 배열로 변

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

환한 뒤 conv-maxpool 노드로 이동한다. conv-maxpool 노드는 Convolution layer와 Pooling layer가 결합된 구조로 되어 있으며, 각각 3개의 layer로 구성되어 있다. dropout 노드에서는 결과값을 도출하기 전에 더욱 좋은 결과를 내기 위해 거치는 과정으로 CNN에서 정규화하는 가장 인기 있는 방법이다. dropout은 설정한 값을 통해 일부의 연산을 사용하지 않도록 한다. 모든 연산이나 값을 사용하는 것 보다 일부의 값을 사용했을 경우 더 좋은 효과를 도출할 수 있다는 가정으로 적용하는 방식이기 때문에 학습을 통해 conv-maxpool 노드에서 도출된 다양한 값을 모두 사용하지 않고 최적의 값을 적용하기 위해 일부의 연산만으로 결과를 나타낸다. 학습하는 동안에는 dropout을 0.5로 설정하고 분류하는 동안에는 1로 dropout을 설정한다. 최종적으로 결과 값을 도출하기 위해 loss 노드와 accuracy 노드를 통해 손실된 값과 정확률을 계산한다.

```
Evaluation:
2016-10-31T12:16:52.620628: step 660, loss 2.42953, acc 0.278

Saved model checkpoint to /home/hong/git/cnn-text-classification-tf/runs/1477883271/checkpoints/model-660

2016-10-31T12:16:53.609488: step 661, loss 2.77837, acc 0.15
2016-10-31T12:16:54.394091: step 662, loss 2.91881, acc 0.05
2016-10-31T12:16:55.173485: step 663, loss 2.95005, acc 0.15
2016-10-31T12:16:55.949323: step 664, loss 2.49881, acc 0.2
2016-10-31T12:16:56.737964: step 665, loss 2.42411, acc 0.15
2016-10-31T12:16:57.512295: step 666, loss 2.25318, acc 0.3
2016-10-31T12:16:58.287738: step 667, loss 2.49773, acc 0.25
2016-10-31T12:16:59.067748: step 668, loss 2.54982, acc 0.3
2016-10-31T12:16:59.844896: step 669, loss 2.93253, acc 0.1
2016-10-31T12:17:00.633005: step 670, loss 2.47822, acc 0.3
2016-10-31T12:17:01.419038: step 671, loss 2.41283, acc 0.25
2016-10-31T12:17:02.196254: step 672, loss 2.54989, acc 0.2
2016-10-31T12:17:02.973565: step 673, loss 2.99131, acc 0.2
2016-10-31T12:17:03.759813: step 674, loss 2.71901, acc 0.1
2016-10-31T12:17:04.569156: step 675, loss 2.5877, acc 0.15
```

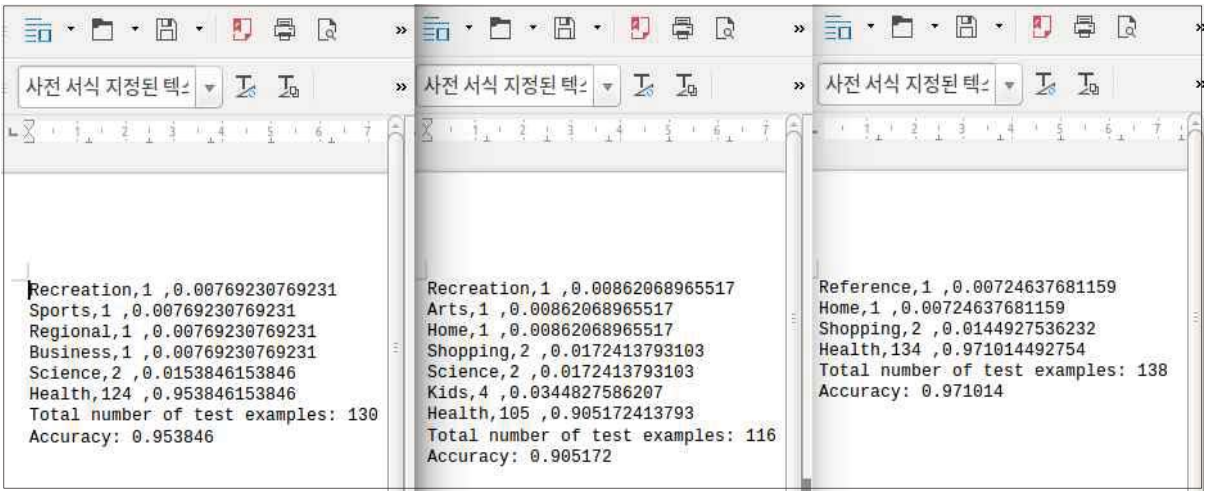
[그림 3-10] 텍스트 분류를 위한 CNN 모델의 학습 내용

[그림 3-10]은 텍스트 분류를 위한 CNN 모델의 학습 내용의 일부를 나타낸다. step을 통해 학습한 단계의 횟수를 알 수 있으며, loss를 통해 학습하는 과정에서 발생하는 손실률에 대해 알 수 있고 acc를 통해 학습이 얼마나 정확하게 되고 있는지를 나타내는 정확률을 알 수 있다. 학습된 모델에 정보제공자에 해당하는 사용자의 텍스트를 입력하여 관심사 카테고리를 기준으로 관심사를 분류했으며, 카테고리마다 10명의 정보제공자를 선정했고 각 정보제공자의 20개의 게시글에서 텍스트

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

를 수집하여 총 3,200 문장을 분류했다. 총 학습한 step 수는 28,440번이다.



[그림 3-11] 정보제공자 텍스트에서 관심사 분류 결과의 예

[그림 3-11]은 세 명의 정보제공자 SNS 게시글에서 텍스트의 문장을 수집하고 학습 모델을 이용하여 분류한 텍스트 결과의 예를 보여준다. 분류된 텍스트의 관심사 카테고리와 빈도수, score를 오름차순 정렬하여 차례로 보여준다. 가장 왼쪽에 있는 텍스트 문서에서는 총 130개의 문장이 등장했으며, 분류한 결과 Health로 분류된 문장이 124개이고 science로 분류된 문장이 2개이다. 나머지 문장은 빈도 1로 Business, Regional, Sports, Recreation으로 분류되었다. 빈도수 옆에 score 값은 총 단어의 개수에서 분류된 단어의 score를 나타낸다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

3. 관심사 카테고리를 이용한 데이터베이스 구축

본 절에서는 앞 절에서 분류한 이미지와 텍스트의 관심사 카테고리를 이용하여 정보제공자 데이터베이스를 구축한다. 정보제공자의 이미지와 텍스트의 관심사 카테고리가 일치할 경우에만 정보제공자 데이터베이스에 정보제공자의 기본적인 정보를 태깅하여 저장한다. 각 관심사 카테고리명을 ID 값으로 정의하여 정보제공자에게 부여하고 이미지와 텍스트의 score 값을 차례로 태깅하는 형태로 저장하여 정보제공자의 이미지와 텍스트의 값을 표현함으로써 어느 정도 수준의 값을 갖는지 파악하는데 용이하게 배치하였다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-2] 정보제공자 데이터베이스

Category	Contents	
Arts	Arts_2(0.1588, 0.6603) Arts_4(0.1609, 0.8915) Arts_7(0.7509, 0.8752)	Arts_3(0.7192, 0.6121) Arts_5(0.7069, 0.5983) Arts_9(0.6500, 0.9997)
Business	Business_1(0.7566, 0.9583) Business_4(0.4579, 0.8012) Business_6(0.3360, 0.5913) Business_9(0.2927, 0.7542)	Business_3(0.4195, 0.4579) Business_5(0.2619, 0.7626) Business_8(0.5959, 0.7167) Business_10(0.1781, 0.7648)
Computers	Computers_1(0.3661, 0.4703) Computers_5(0.2946, 0.9085)	Computers_3(0.0868, 0.2751) Computers_8(0.1709, 0.8974)
⋮	⋮	⋮
Reference	Reference_1(0.3483, 0.8101) Reference_4(0.4513, 0.9548) Reference_7(0.5481, 0.9561)	Reference_2(0.6558, 0.8658) Reference_6(0.5252, 0.9184) Reference_9(0.0776, 0.6894)
Regional	Regional_2(0.6403, 0.9267) Regional_5(0.6492, 0.8678) Regional_8(0.4385, 0.8901)	Regional_3(0.4495, 0.8502) Regional_7(0.2687, 0.7292) Regional_10(0.3942, 0.6131)
Science	Science_1(0.4776, 0.7738) Science_3(0.6110, 0.7029) Science_6(0.2383, 0.7633) Science_8(0.6309, 0.8412)	Science_2(0.6420, 0.9811) Science_4(0.4934, 0.7216) Science_7(0.5806, 0.7374) Science_9(0.6171, 0.7659)
Shopping	Shopping_1(0.4248, 0.6183) Shopping_3(0.5879, 0.8054) Shopping_6(0.4716, 0.9614) Shopping_8(0.6997, 0.5612)	Shopping_2(0.5891, 0.8248) Shopping_5(0.5017, 0.7194) Shopping_7(0.4709, 0.6298) Shopping_10(0.3640, 0.8326)
Society	Society_2(0.4427, 0.7823) Society_7(0.3186, 0.8665)	Society_5(0.5872, 0.9982) Society_10(0.5136, 0.9586)
Sports	Sports_1(0.6074, 0.7860) Sports_8(0.4299, 0.7041)	Sports_5(0.4108, 0.9494) Sports_10(0.6362, 0.8741)
Kids	Kids_1(0.3816, 0.8066) Kids_3(0.5665, 0.8703) Kids_6(0.5964, 0.9187) Kids_8(0.5180, 0.6500)	Kids_2(0.3404, 0.7435) Kids_4(0.4612, 0.7200) Kids_7(0.3965, 0.6741) Kids_9(0.5760, 0.6680)
World	World_1(0.4366, 0.8695) World_6(0.3096, 0.9777)	World_5(0.3952, 0.9076) World_10(0.8604, 0.9236)

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-2]는 정보제공자 데이터베이스를 나타낸다. 관심사가 Arts로 분류되어 Arts 카테고리에 해당하는 정보제공자에게는 Arts_n의 ID를 부여하며, n은 순서를 나타낸다. ID의 소괄호 안에 숫자는 이미지의 score값과 텍스트의 score값을 차례대로 태깅한 것이다. Arts 카테고리의 Arts_2 정보제공자의 이미지와 텍스트의 score 값은 0.1588, 0.6603이고 Business 카테고리의 Business_1 정보제공자의 이미지와 텍스트의 score 값은 0.7566, 0.9583 인 것을 알 수 있다. 정보제공자 ID 값이 1부터 시작하지 않는 경우나 숫자가 순차적이지 않고 비어있는 경우는 분류한 정보제공자의 관심사 카테고리가 일치하지 않는 경우이다.

사용자에게 관심사 카테고리 기반의 팔로잉 추천을 제공하기 위해 정보제공자 데이터베이스를 이용한다. 사용자의 SNS 게시글에서 이미지와 텍스트를 추출 한 뒤 관심사 카테고리 기반으로 분류하고 같은 관심사를 갖는 정보제공자를 추천해 주며, 이미지와 텍스트의 score 값을 기반으로 가장 유사한 관심사를 갖는 정보제공자를 추천한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. 사용자 팔로잉 추천

본 절에서는 사용자에게 팔로잉을 추천해주기 위한 방법에 관해 기술한다. 3절에서 정보제공자의 관심사 카테고리를 분류한 방법과 동일하게 사용자의 관심사 카테고리를 분류하고 이미지와 텍스트 값을 도출하며, 정보제공자 데이터베이스를 이용하여 사용자와 같은 관심사 카테고리를 갖는 정보제공자 중에 가장 유사한 관심사를 갖는 정보제공자를 추천한다.

1. 사용자 관심사 카테고리 분류

사용자 SNS 게시글의 이미지와 텍스트를 이용하여 관심사 카테고리를 분류하는 방법에 관해 기술한다. 3절에서 기술한 정보제공자의 관심사를 분류하는 방법과 동일한 방법을 이용하여 사용자의 관심사 카테고리를 분류한다. 학습된 이미지와 텍스트 CNN 모델에 사용자의 이미지와 텍스트를 입력하여 관심사 카테고리를 분류한다. 관심사를 분류하기 위해 입력하는 이미지와 텍스트의 갯수는 이미지 20개와 20개의 게시글에서 추출한 텍스트이며, 총 144명의 사용자를 대상으로 관심사 카테고리를 분류한 후 정보제공자에게 ID를 부여하고 score 값을 태깅한 것과 마찬가지로 사용자에게도 ID를 부여하고 score 값을 태깅한다. 다음은 사용자의 관심사를 분류한 결과이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-3] 사용자 관심사 분류 테이블

ID	Category	Score
User_1	Arts	0.3856, 0.9538
User_2	Business	0.4303, 0.9747
User_3	Computers	0.5597, 0.9462
User_4	Games	0.4321, 0.9548
User_5	Health	0.5564, 0.9600
User_6	Home	0.4843, 0.9000
User_7	News	0.6165, 0.6923
User_8	Recreation	0.5093, 0.9432
User_9	Reference	0.5362, 0.8621
User_10	Regional	0.5266, 0.9680
User_11	Science	0.5445, 0.9808
User_12	Shopping	0.5442, 0.9538
User_13	-	-
User_14	Sports	0.4850, 0.9432
User_15	Kids	0.4729, 0.9701
⋮	⋮	⋮
User_136	-	-
User_137	Reference	0.3774, 0.9800
User_138	Regional	0.3830, 0.9502
User_139	Science	0.2544, 0.9737
User_140	Shopping	0.5426, 0.9358
User_141	Society	0.5103, 0.9640
User_142	Sports	0.5098, 0.9640
User_143	Kids	0.4269, 0.9879
User_144	-	-

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

144명의 사용자에 대해 User_n 형식의 ID를 부여하였으며, n에는 숫자의 값을 통해 사용자를 구분하였다. User_1의 경우에 Category는 Arts로 분류되었고 이미지와 텍스트 값에 해당하는 score 값은 0.3856, 0.9538이 태깅되었다. Category와 Score 값에 '-'으로 표시된 부분은 분류된 이미지와 텍스트의 카테고리가 일치하지 않은 것을 뜻한다.

일차적으로 정보제공자 데이터베이스의 정보제공자의 관심사 카테고리와 사용자의 관심사 카테고리와 비교하여 같은 카테고리에 해당하는 정보제공자들을 추천하며, 가장 유사한 정보제공자를 팔로잉 추천하기 위한 방법은 다음 절에서 설명한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 거리 측정을 통한 팔로잉 추천

사용자와 가장 유사한 내용을 갖는 정보제공자를 팔로잉 추천하기 위해 사용자와 정보제공자 간 거리를 측정한다. score 값을 x와 y축을 갖는 좌표평면의 점으로 정의하고 같은 카테고리를 갖는 정보제공자와 사용자 사이의 거리를 유클리디안 거리를 이용하여 측정하며, 다음과 같은 공식으로 나타낼 수 있다[42].

$$Distance = \sqrt{\sum_{i=1}^{n}(p^i - q^i)^2} \qquad (2)$$

두 점 P와 Q가 각각 $P=(p_1, p_2, p_3, \ldots, p_n)$와 $Q=(q_1, q_2, q_3, \ldots, q_n)$의 좌표 값을 가질 때 식 (2)를 통해 두 점 사이의 거리를 계산 할 수 있다.

유클리디안 거리 측정법은 개체의 속성 값이 여러개인 경우 속성 값에 따른 개체의 유사도를 구할 때 자주 사용되는 방법으로 속성 값의 개수에 따라 다차원의 유클리드 공간을 정의할 수 있으며, 거리 측정이 가능하다.

같은 관심사 카테고리를 갖는 정보제공자와 사용자 중에서도 가장 유사한 관심사의 관계를 가질 때 팔로잉 추천을 해주기 위해 유사도 거리 측정을 하며, [표 3-4]와 [표 3-5]는 관심사 카테고리가 Arts로 같은 정보제공자와 사용자의 이미지와 텍스트를 보여준다.

동일한 관심사 카테고리 내의 정보제공자와 사용자의 유사도를 측정하기 위한 식 (2)를 이용하여 정보제공자와 사용자 사이의 유사도 거리를 측정한 결과는 [표 3-6]과 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-4] 관심사 카테고리가 Arts인 정보제공자와 사용자의 이미지 예

Interest Category - Arts			
Information Provider ID	Contents	User ID	Contents
Arts_2		User_1	
		⋮	⋮
		User_17	
⋮	⋮	⋮	⋮

[표 3-5] 관심사 카테고리가 Arts인 정보제공자와 사용자의 텍스트 예

Interest Category - Arts			
Information Provider ID	Contents	User ID	Contents
Arts_2	Can you please do a Michael Jackson thriller cake	User_1	Wow even MetDaan is loving this cake! 11 million views already <3
	Check out this week's Harry Potter Cake!		⋮
			"Learn to Fly" mixed media illustration.
	⋮	⋮	⋮
	@walkingdeadamc premier tonight!!!!	User_17	#JosefKote #DirectInsight #FineArt #Private #Collection
	1) What weight of fondant did you use?		⋮
			Good morning Happy Wednesday to all
⋮	⋮	⋮	⋮

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

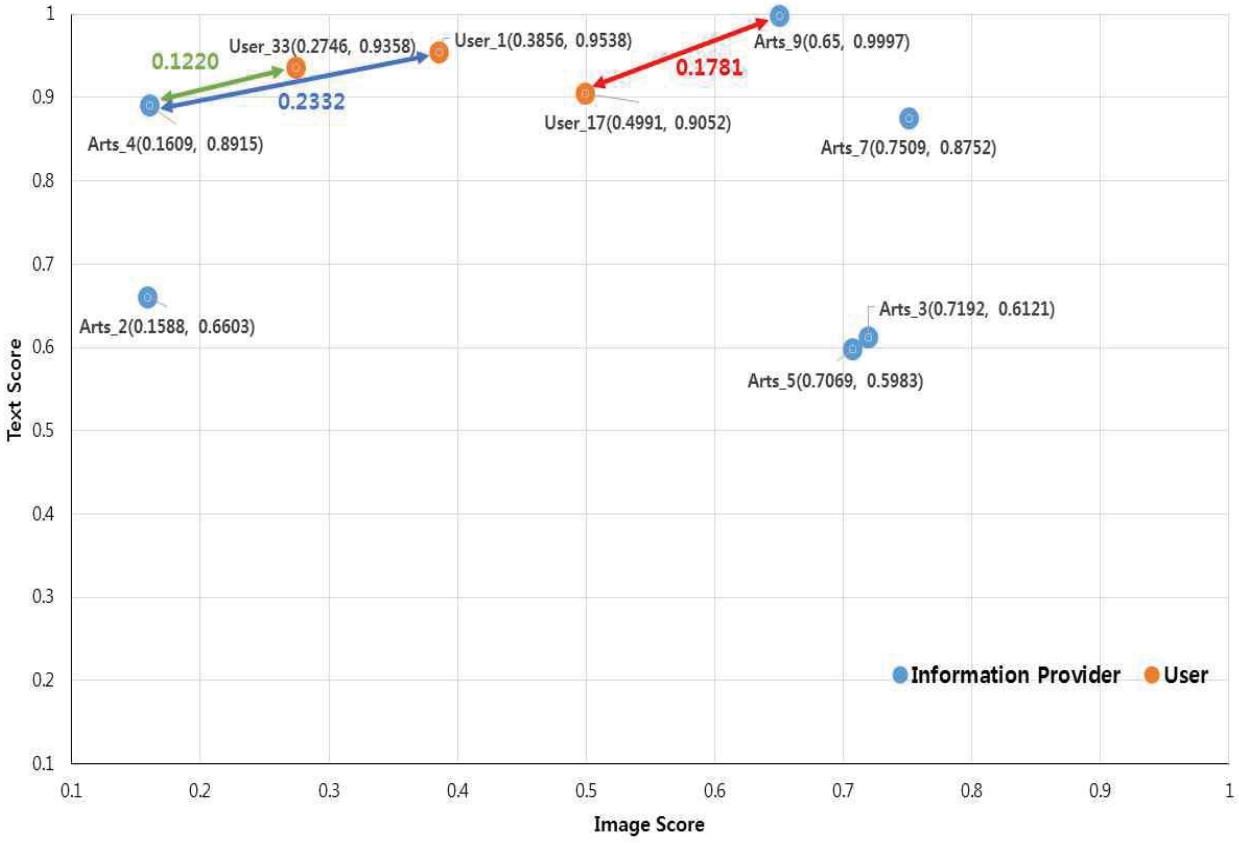
[표 3-6] 정보제공자와 사용자간 거리 유사도 측정 테이블

Category	Information Provider ID	User ID	Distance Value
Arts	Arts_2	User_1	0.3709
		User_17	0.4192
		User_33	0.2988
		⋮	⋮
	Arts_3	User_1	0.4775
		User_17	0.3665
		User_33	0.5500
		⋮	⋮
	Arts_4	User_1	0.2332
		User_17	0.3385
		User_33	0.1220
		⋮	⋮
	Arts_5	User_1	0.3737
		User_17	0.3706
		User_33	0.5484
		⋮	⋮
	Arts_7	User_1	0.4801
		User_17	0.2535
		User_33	0.4801
		⋮	⋮
	Arts_9	User_1	0.2683
		User_17	0.1781
		User_33	0.3808
		⋮	⋮
⋮	⋮	⋮	⋮
World	World_5	User_96	0.1231
		User_112	0.3124
		–	–
		⋮	⋮
	World_6	User_96	0.2012
		User_112	0.4045
		–	–
		⋮	⋮
	World_10	User_96	0.2040
		User_112	0.1539
		–	–
		⋮	⋮

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-4]는 식 (2)를 이용하여 정보제공자 데이터베이스 내의 정보제공자의 이미지와 텍스트의 태깅 값과 사용자의 태깅 값을 계산한 거리 값이다. Arts 카테고리에 존재하는 정보제공자는 Arts_1부터 Arts_7까지 6명이며, Arts 카테고리에 존재하는 사용자는 User_1, User_17, User_33으로 총 3개이다. 모든 정보제공자와 모든 사용사 사이의 유사도 거리를 측정하여 가장 가까운 거리로 결과가 도출되는 경우 팔로잉 추천을 한다. Arts_4 정보제공자와 User_1, User_33 사용자를 팔로잉 추천해 줄 수 있으며, Arts_9 정보제공자와 User_17 사용자를 팔로잉 추천 할 수 있다. [그림 3-12]는 거리 유사도를 측정하여 팔로잉 추천이 가능한 경우를 도식화 한 것이다.



[그림 3-12] 정보제공자와 사용자 사이의 거리 유사도 측정 결과

사용자의 관심사를 분류하여 관심사를 통한 팔로잉 추천이 가능하지만 같은 관심사 내에서도 사용자의 성향이나 세부적인 관심사가 다를 수 있다. 따라서 본 논문에서는 사용자 사이의 거리를 측정하여 가장 유사한 정보를 가진 정보제공자를 팔로잉 추천하기 때문에 가장 적합한 팔로잉 추천이 가능하다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 실험 및 결과

본 장에서는 본 연구에서 데이터 학습과 관심사 카테고리 분류를 위해 사용한 실험 환경과 학습과 분류를 위해 수집한 학습 데이터 셋과 실험 데이터 셋에 대해 설명한다.

또한 본 논문에서 제안한 관심사 카테고리를 기반으로 이미지와 텍스트를 함께 사용하여 모델링한 CNN으로 관심사를 분류하는 방법이 얼마나 효과적인지 검증하기 위해 기존에 제안된 방법과 비교 및 실험을 한다. 텍스트를 이용한 경우는 빈도수를 이용한 경우와 LDA를 이용한 경우를 비교하였으며, 텍스트와 이미지를 함께 사용했지만 텍스트는 빈도수를 이용하고 관심사를 기준으로 태깅되지 않은 CNN을 이용하여 이미지를 분류하는 방법을 비교한다[9,10].

본 연구에서 실험에 사용한 환경은 다음과 같다.

[표 4-1] 실험 환경

Division		Contents
HW	CPU	Intel(R) Core(TM) i7-4712MQ CPU @ 2.30GHZ
	GPU	NVDIA GeForce 840M
	RAM	8GB
	HDD	200GB
SW	OS	Windos7, Ubuntu16.04 LTS
	Program Language	Java, Python

A. 데이터 수집

본 절에서는 본 연구에서 사용한 SNS 게시글의 이미지와 텍스트를 수집하는 방법에 대해 기술한다.

이미지는 인스타그램에서 수집하였으나 Instagram Developer에서 제공하는 API를 사용하기 위해서는 많은 제약 사항이 존재하기 때문에 크롬 확장프로그램인 Fatkun Batch Download Image를 이용하여 이미지를 수집하였다[38].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

학습하기 위한 텍스트와 관심사 분류를 위한 정보제공자와 사용자의 이미지와 텍스트는 Java기반의 Facebook API인 Facebook4j를 이용하여 수집하였다[41].

개발 환경으로는 JAVA JDK 1.8 버전을 사용하였고, Apache Maven 3.3.1 버전과 Eclipse neon을 사용했다.

[표 4-2] 페이스북에서 텍스트와 이미지 수집

```
Reading reading = new Reading().limit(25);
ResponseList<Post> feeds = facebook.getFeed(userId, reading);

for (int i = 0; i < feeds.size(); i++) {
        if (feeds.size() < 25) break;
            Post post = feeds.get(i);
            int num = i + 1;
            String message = post.getMessage();
            URL picture = post.getPicture();
            if (message != null) {
                    if (message.contains("\n")) {
                        String[] m = message.split("\n");
                        for (int j = 0; j < m.length; j++) {
                        System.out.println(i + 1 + ":::" + m[j] + "\t" + tag);
                        }
                    } else {
                        System.out.println(i + 1 + ":::" + message + "\t" + tag);
                    }
             }
         if (picture != null) {
             System.out.println(num + "::" + picture);
             URL url = new URL(picture.toString());
             BufferedImage img = ImageIO.read(url);
             File file = new File(path + tag + "_" + num + ".jpeg");
             ImageIO.write(img, "jpeg", file);
          }
```

B. 데이터 셋

본 절에서는 본 연구에서 수집한 데이터 셋에 대하여 설명한다. 데이터 셋은 크게 제안한 방법에 사용된 데이터 셋과 기존 방법에 사용된 데이터 셋으로 구성된다.

카테고리 기반의 관심사를 분류하기 위한 학습 데이터 셋의 이미지는 인스타그램에서 수집했고 텍스트는 페이스북에서 수집했다. 제안하는 방법에서 정보제공자

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

와 사용자의 관심사를 분류하기 위한 실험 데이터 셋과 기존에 제안한 방법으로 관심사를 분류하기 위한 실험 데이터 셋은 페이스북에서 수집했으며, [표 4-3]은 본 연구에서 사용한 전체 데이터 셋을 나타낸다.

[표 4-3] 전체 데이터 셋

Division		SNS	Image	Text (Sentence)	Text (Word)
제안한 방법	학습	Instagram	16,000	-	-
		Facebook	-	16,000	-
	정보제공자	Facebook	3,200	3,200	-
	사용자	Facebook	2,880	2,880	-
기존 방법	텍스트 빈도수	Facebook	-	-	5,424
	LDA	Facebook	-	-	6,000
	텍스트 빈도수 + 이미지	Facebook	900	-	5,424

C. 실험 평가 방법 및 결과 분석

1. 실험 평가 방법

본 논문의 실험은 제안한 관심사 카테고리 분류가 얼마나 효과적으로 이루어졌는지 평가하기 위해 정확률을 이용하며, 식 (3)을 통해 측정한다.

$$Precision(\%) = \frac{\text{올바르게 분류한 게시글의 수}}{\text{분류한 게시글의 수}} \times 100 \qquad (3)$$

정확률은 이미지와 텍스트를 이용하여 분류한 관심사 카테고리가 얼마나 정확한지 평가하는 것이며, 분류한 모든 게시글에서 올바르게 분류된 게시글이 얼마나 되

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

는지를 측정한다.

기존의 관심사 분류 방법은 본 논문에서 제안한 방법과 동일하게 관심사 카테고리를 기준으로 DMOZ를 이용하였으며, 텍스트의 빈도수를 이용한 경우와 텍스트에 LDA를 이용한 경우, 텍스트의 빈도수와 관심사 카테고리로 태깅하지 않은 CNN 모델을 이용하여 이미지를 분류한 경우의 정확률을 측정하고 비교하여 본 논문에서 제안한 방법의 성능을 검증한다[9,10]. 게시글에서 텍스트의 빈도수를 이용하여 관심사를 분류하는 경우 전처리 과정을 이용하여 불용어를 제거하고 단어를 추출한 후 빈도수를 측정하여 상위 5개의 단어와 DMOZ를 비교하여 관심사를 분류한다[10]. LDA를 이용하여 관심사를 분류하는 경우 전처리 과정을 이용하여 불용어를 제거하고 LDA 알고리즘을 적용하여 토픽을 분류한 후 DMOZ와 비교하여 관심사를 분류한다[9]. 텍스트의 빈도수와 관심사 카테고리로 태깅되지 않은 CNN 모델을 이용하여 관심사를 분류하는 방법에서 텍스트의 빈도수는 앞서 서술한 방법과 일치하며 기존에 모델링 되어 있는 Incpetion-v3 모델을 수정하지 않고 그래도 이용하여 이미지를 분류 한 후 텍스트의 관심사와 이미지의 관심사를 이용하여 최종적으로 사용자의 관심사를 분류한다[10, 40].

2. 실험 결과 분석

식 (3)에서 분류한 게시글의 수는 분류한 정보제공자와 사용자의 수와 같으며, 올바르게 분류한 게시글의 수는 올바르게 분류된 정보제공자와 사용자의 수와 같다. 이를 이용하여 평가한 관심사 카테고리 분류에 관한 성능은 [표 4-4]와 [표 4-5], [그림 4-1], [그림 4-2], [그림 4-3]과 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-4] 정보제공자 관심사 카테고리 분류 성능

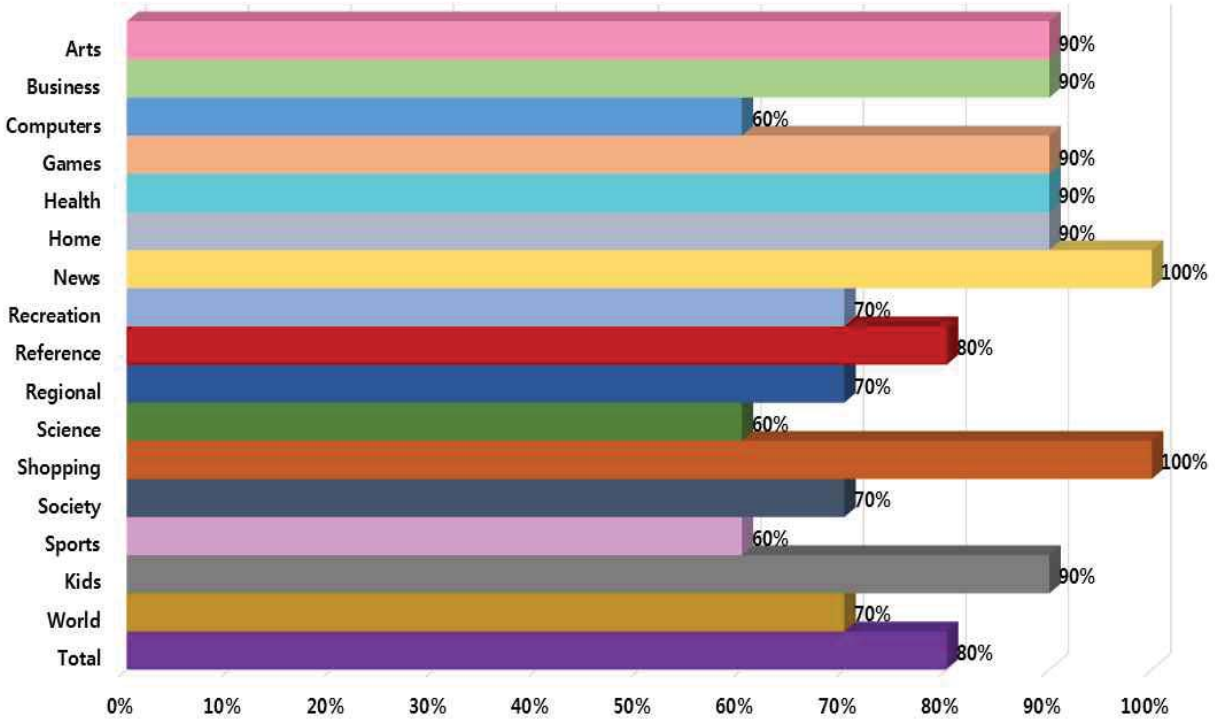
Category	Precision(%)	Category	Precision(%)
Arts	90%	Business	90%
Computers	60%	Games	90%
Health	90%	Home	90%
News	100%	Recreation	70%
Reference	80%	Regional	70%
Science	60%	Shopping	100%
Society	70%	Sports	60%
Kids	90%	World	70%
Total Precision(%)		80%	

[표 4-5] 사용자 관심사 카테고리 분류 성능

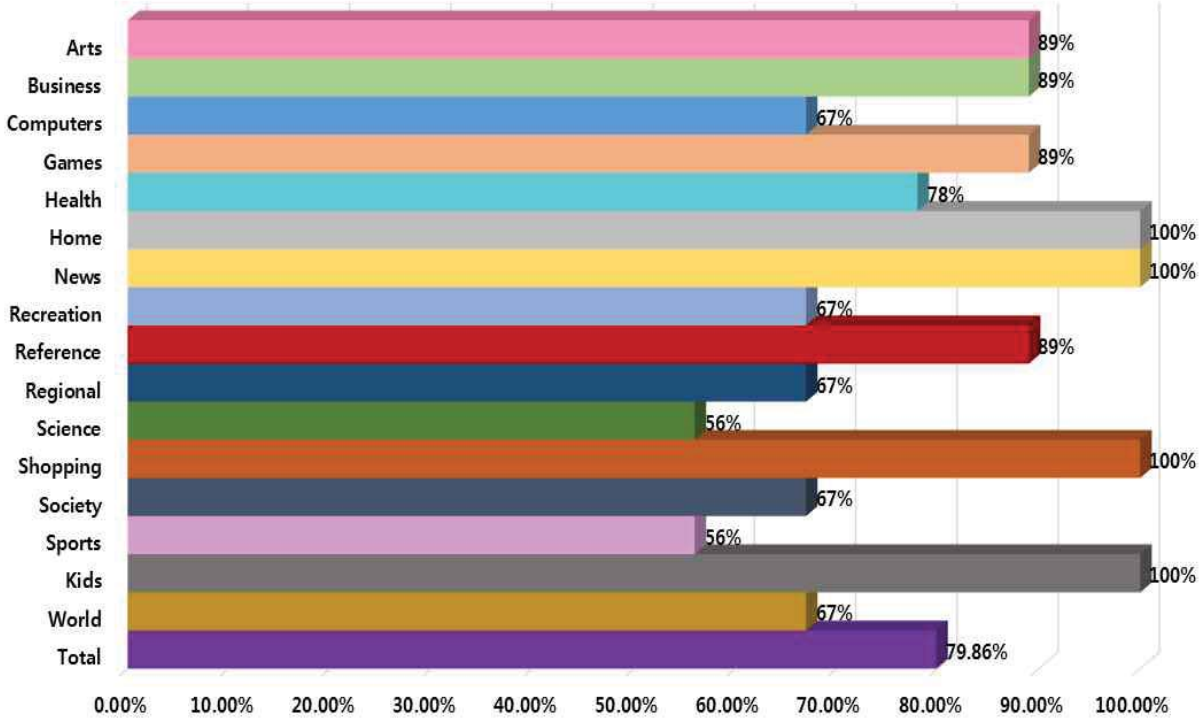
Category	Precision(%)	Category	Precision(%)
Arts	89%	Business	89%
Computers	67%	Games	89%
Health	78%	Home	100%
News	100%	Recreation	67%
Reference	89%	Regional	67%
Science	56%	Shopping	100%
Society	67%	Sports	56%
Kids	100%	World	67%
Total Precision(%)		79.86%	

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 4-1] 정보제공자 카테고리 분류 성능 그래프

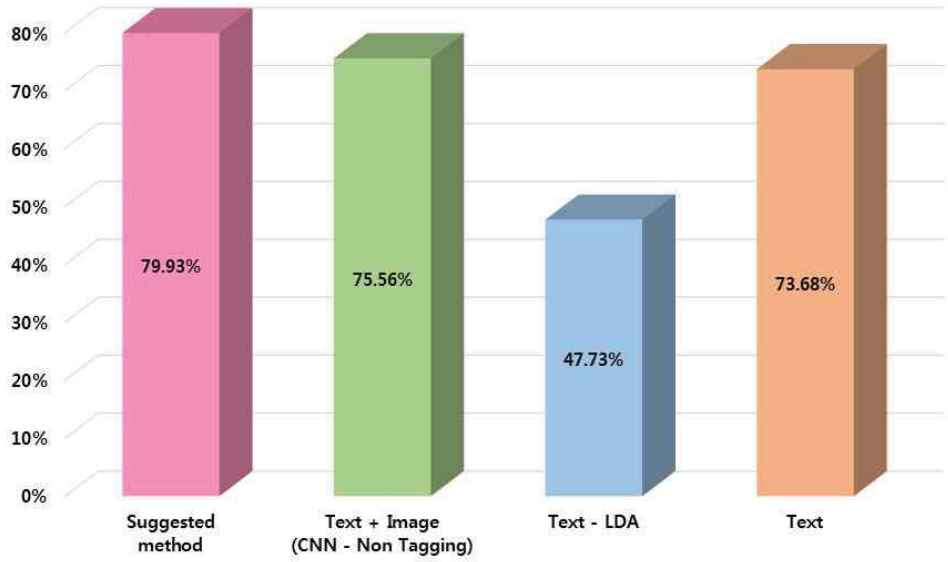


[그림 4-2] 사용자 카테고리 분류 성능 그래프

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-4]와 [표 4-5], [그림 4-1], [그림 4-2]를 통해 본 논문에서 제안한 사용자 관심사 카테고리 분류 방법에 대한 성능을 보여준다. [표 4-4]와 [그림 4-1]은 정보 제공자의 관심사 카테고리 분류에 대한 정확률을 나타내고 있으며, 전체 80%의 정확률을 보인다. [표 4-5]와 [그림 4-2]는 사용자의 관심사 카테고리 분류에 대한 정확률을 나타내며, 전체 79.86%의 정확률을 보인다.



[그림 4-3] 기존 방법과 제안하는 방법에 대한 성능 비교

[그림 4-3]은 제안하는 방법과 기존의 관심사 분류 방법에 대한 정확률을 나타내고 있다. 기존의 방법 모두 16개의 DMOZ 카테고리를 관심사로 분류하며, [9,10]에서 제안하는 방법을 이용하여 실험했다. SNS 게시글 내의 텍스트 빈도수만 이용하였을 경우 73.33%의 정확률을 보이고 문서에서 토픽을 추출하는 알고리즘인 LDA를 이용할 경우 47.73%의 정확률, 텍스트의 빈도수와 태깅할 카테고리를 이용하지 않은 CNN을 이용할 경우 75.56%의 정확률을 보였으나 제안한 방법을 이용할 경우 79.93%의 정확률을 보인다.

[그림 4-1], [그림 4-2]를 통해 정보제공자와 사용자 카테고리 분류가 전반적으로 잘 되었음을 알 수 있으며, Kids, Shopping, News, Home 카테고리의 경우 높은 정확률을 보이는데 이는 사용자가 대부분 유사한 이미지와 텍스트 데이터를 게시하기 때문인 것으로 보인다. Sports, Science 카테고리의 경우 다른 카테고리에 비해 낮은 정확률을 보인다. Sports의 경우에는 Recreation과 비슷한 유형의 데이터

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

가 다수 포함되어 있어서 좋은 성능을 보이지 못했고 Science와 Computers 역시 비슷한 유형의 데이터가 다수 포함되어 있었으며, Regional과 World, Games와 Arts 역시 비슷한 유형의 데이터가 다수 포함되어 최상의 성능을 보이지 못했다. 관심사 카테고리에서 뜻이 비슷하거나 내포되는 단어는 이미지와 텍스트 데이터가 유사하거나 겹치는 현상이 발생하였으며, 카테고리 사이에 상관관계가 있다는 것을 알 수 있다. 따라서 SNS 사용자의 관심사를 정확하게 분류하기 위해서는 SNS에 알맞은 관심사 카테고리의 기준을 정의하는 연구가 필요할 것으로 보인다. 또한 [그림 4-3]을 통해 제안하는 방법이 기존의 관심사 분류 방법이 더 뛰어난 성능을 보임을 알 수 있다.

기존의 SNS 팔로잉 추천은 관계 중심이며, SNS 게시글에서 이미지와 텍스트 중 하나만을 이용하지만 본 연구에서는 게시글의 이미지와 텍스트를 함께 사용하는 정보중심의 관심사 카테고리 분류 방법을 제안함으로써 정확한 관심사 분류를 기반으로 한 팔로잉 추천을 통해 다양한 정보를 습득할 수 있는 양질의 서비스를 제공 할 수 있을 것으로 보이며, 관심사 기반의 사용자 맞춤형 서비스, SNS 마케팅 서비스 등의 분야에서 활용 될 수 있을 것으로 기대된다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

V. 결론 및 제언

본 논문에서는 SNS 게시글에서 이미지와 텍스트를 수집하여 CNN을 통해 학습한 후 사용자의 관심사를 카테고리별로 분류하고 많은 정보를 보유하고 있는 정보제공자를 팔로잉 추천해주는 방법을 제안했다.

기존의 SNS 관련 연구와 팔로잉 추천을 위한 연구에서는 이미지와 텍스트를 각각 사용하였기 때문에 사용자의 의도를 정확히 파악하기 어려운 점이 있었으며, 관계 중심으로 팔로잉 추천을 제공하므로 정보공유 및 습득을 원하는 사용자에게 적합하지 않았다. 따라서 본 연구에서는 게시글에 존재하는 이미지와 텍스트를 함께 사용하여 사용자의 의도를 정확히 파악 할 수 있으며, 관심사 분류의 명확한 기준을 정의하기 위해 사람이 직접 참여하는 ODP로 구축되는 DMOZ를 관심사 카테고리로 이용함으로써 객관성을 높이고 기계학습 중 하나인 CNN을 이용하여 이미지와 텍스트를 학습 및 분류한 후 같은 카테고리 내에 존재하는 사용자와 정보제공자 사이의 거리를 측정하여 가장 유사한 관심사를 갖는 정보제공자를 팔로잉 추천하는 방법을 제안한다. 제안한 방법의 실험 결과 16개에 해당하는 전체 관심사 카테고리 분류에 대한 정보제공자의 정확률은 80%를 보였고 사용자의 정확률은 79.86%를 보임으로써 전체 79.93%의 정확률을 보였다. 비교적 높은 정확률을 보인 카테고리는 Kids, Shopping, News, Home 이였으며, 낮은 정확률을 보인 카테고리의 경우 카테고리의 뜻이 비슷하거나 내포되는 의미를 가짐으로써 유사한 이미지와 텍스트 데이터가 다수 포함되기 때문으로 분석된다.

본 연구는 DMOZ를 이용하여 관심사 카테고리로 분류하였으나 중첩되는 뜻이나 의미를 지니지 않는 완전히 SNS에 적합한 관심사 카테고리를 정의하여 이미지와 텍스트 데이터를 이용하여 관심사를 분류한다면 좀 더 명확한 결과가 도출 될 것으로 보이며, 향후 연구로 SNS의 이미지를 학습하고 분류하기에 적합한 깊이를 갖는 CNN 모델 구축과 SNS의 특성 상 단문과 특수문자, 이모티콘으로 구성되어 있는 텍스트를 학습하고 분류하기에 적합한 CNN 모델 구축에 관한 연구가 필요하다.

제안하는 방법을 통해 사용자의 관심사에 적합한 정보를 추천해주는 사용자 맞춤형 서비스, SNS 마케팅 등의 분야에서 활용 할 수 있을 것으로 기대된다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

참고문헌

[1] 노연우, 김대윤, 한지은, 육미선, 임종태, 복경수, 유재수, “소셜 네트워크에서 사용자의 영향력을 고려한 핫 토픽 예측 기법”, 한국콘텐츠학회논문지, 제 15권, 제 8호, pp. 24-36, 2015.

[2] 김윤화, “SNS 이용추이 및 이용행태 분석,” KISDI STAT Report, KISDI STAT Report 16-07, pp. 1-9, 2016.

[3] 박병선, “글로벌 SNS 이용 현황과 시사점,” 정보통신방송정책, 제 26권, 제 2호 통권 570호, pp. 22-34, 2014.

[4] 차경진, 이은목, “사회비교이론 관점에서 살펴본 SNS 이용중단 의도”, 한국전자거래학회지, 제 20권, 제 3호, pp. 59-77, 2015.

[5] 홍명덕, 오경진, 가명현, 조근식, “개인화된 뉴스 서비스를 위한 소셜 네트워크 기반의 콘텐츠 추천 기법”, 제 19권, 제 3호, pp. 57-71, 2013.

[6] 전련화, 김연정, 김보현, 이민수, “소셜 네트워크 기반의 개인 맞춤형 정보추천 기법”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제 19권, 제 12호, pp. 668-672, 2013.

[7] 유소엽, 정옥란, “소셜 카테고리를 이용한 추천 방법”, 인터넷정보학회논문지, 제 15권, 제 5호, pp. 73-82, 2014.

[8] 김경민, 김동윤, 이지형, “트위터를 활용한 감정 기반의 영화 유사도 측정”, 한국지승시스템학회 논문지, 제 24권, 제 3호, pp. 292-297, 2014.

[9] 홍택은, 주영지, 신주현, “토픽과 카테고리를 이용한 SNS 팔로잉 추천 방법”, 한국멀티미디어학회 춘계학술발표대회, 제 19권, 제 1호, pp. 89-92, 2016.

[10] 홍택은, 신주현, “이미지와 텍스트 정보의 카테고리 분류에 의한 SNS 팔로잉 추천 방법”, 스마트미디어저널, 제 5권, 제 3호, pp. 54-61, 2016.

[11] Statista, https://www.statista.com/

[12] Kantar TNS Global, http://www.tns-infratest.com/

[13] D. H. Hubel and T.N Wiesel, “Receptive fields of single neurones in the cat's striate cortex”, Journal of Physiology, Vol, 143, No. 3, pp. 574-591, 1959.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[14] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat′s visual cortex”, Journal of Physiology, Vol. 160, No. 1, pp. 106-154, 1962.

[15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, “Backpropagation applied to handwritten zip code recognition”, Neural Computaion, Vol, 1, Issue 4, pp. 541-551, 1989.

[16] S. Behnke, “Hierarchical Neural Networks for Image Interpretation”, Springer-Verlag Berlin Heidelberg, 2003.

[17] P. Y. Simard, D. Steinkraus, J. C. Platt, “Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis”, ICDAR, Vol. 2, pp. 958, 2003.

[18] 안성만, “딥러닝의 모형과 응용사례”, 지능정보연구, 제 22권, 제 2호, pp. 127-142, 2016.

[19] 장현웅, 조수선, “CNN을 이용한 소셜 이미지 자동 태깅”, 정보과학회논문지, 제 43권, 제 1호, pp. 47-53, 2016.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause and et al., “ImageNet Large Scale Visual Recognition Challenge”, International Journal of Compuer Vision, No. 115, Vol. 3, pp. 211-252, 2015.

[21] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems 25, 2014.

[22] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks”, ECCV, 2014.

[23] C. Szegedy, W. Liu, Y. Jia and et al., “Going deeper with convolutions”, Computer Vision and Pattern Recognition, 2015.

[24] K. He, X. Zhang, S. Ren and et al., “Deep Residual Learning for Image Recognition”, Computer Vision and Pattern Recognition, 2015.

[25] Y. Kim, “Convolutional Neural Networks for Sentence Classification”, EMNLP, 2014.

[26] N. Kalchbrenner, E. Grefenstette, P. Blunsom, “A Convolutional Neural Network for Modelling Sentences”, ACL, 2014.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[27] C. N. Santos, M. Gatti, “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts”, COLING, 2014.

[28] R. Johnson, T. Zhang, “Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding”, NIPS, 2015.

[29] WILDML, http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/

[30] cnn-text-classification-github, https://github.com/dennybritz/cnn-text-classification-tf

[31] M. Abadi, A. Agarwal, P. Barham and et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”, Distributed, Parrel, and Cluster Computing, 2016.

[32] M. Abadi, P. Barham, J. Chen and et al., “TensorFlow: A system for large-scale machine learning”, Distributed, Parrel, and Cluster Computing, 2016.

[33] tensorflow, https://www.tensorflow.org/

[34] DOMZ, http://www.dmoz.org/

[35] 남민지, 김정인, 신주현, “인스타그램 기반 이미지와 텍스트를 활용한 사용자 감정정보 측정”, 한국멀티미디어학회, 제 17권, 제 9호, pp. 1125-1133, 2014.

[36] 김혜빈, “인스타그램 이용자 성격이 브랜드 만족도와 충성도에 끼치는 영향”, 한국콘텐츠학회논문지, 제 16권, 제 6호“, pp. 450-431, 2016.

[37] instagram developer, https://www.instagram.com/developer/

[38] Fatkun, https://chrome.google.com/webstore/detail/fatkun-batch-download-ima/nnjjahlikiabnchcpehcpkdeckfgnohf?hl=ko

[39] C. Szegedy, V. Vanhoucke, S. loffe and et al., “Rethinking the Inception Architecture for Computer Vision”, Computer Vision and Pattern Recognition, 2015.

[40] tensorflow-inception, Serving Inception Model with TensorFlow Serving and Kubernetes

[41] facebook4j, http://facebook4j.github.io/en/index.html

[42] 윤종찬, 김학철, 김종진, 윤성대, “유사도와 유클리디안 계산패턴을 이용한 CBR 패턴연구”, 한국정보통신학회논문지, 제 14권, 제 4호, pp. 875-885, 2010.

Collection @ chosun