논문 19-44-09-18 The Journal of Korean Institute of Communications and Information Sciences '19-09 Vol.44 No.09
https://doi.org/10.7840/kics.2019.44.9.1742

SNS에서 텍스트-칼라영상의 교차양식 검색을 위한 데이터셋 구축과 딥러닝 모델

김 강 섭◆, 이 준 환°

Dataset Construction and Deep Learning Models for Cross-Modal Retrieval for Text-Color Image in SNS

KangSub Kim◆, Joonwhoan Lee°

요 약

ImageNet 등과 같은 대규모 영상 데이터셋 구축이 컴퓨터 비전기술 발전에 많은 기여를 하듯이, 최근 여러 종류의 대규모 데이터셋의 구축은 해당 양식을 이용하는 인공지능 분야 발전에 큰 도움이 되고 있다. 그러나, 한국어 텍스트와 칼라영상의 교차양식 검색을 위한 대규모의 한국어 데이터셋은 존재하지 않는다. 논문에서는 소셜 네트워크 서비스(SNS) 중 하나인 인스타그램 게시물인 텍스트와 칼라영상을 수집, 정리하여 텍스트와 칼라영상의 교차양식 검색을 위한 대규모 데이터셋 구성방법을 제안하였다. 또한 데이터셋의 활용 가능성을 보이기 위해 주위기반의 딥러닝 모델 등 서로 다른 두 양식사이의 상호간의 교차양식 검색 방법들을 적용하고, 그 성능을 평가하였다. 본 연구에서 구성된 데이터셋은 SNS 등의 짧은 길이 한국어 문장 분석을 필요로 하는 다양식 기계학습에 활용할 수 있으며, 교차양식 검색 성능평가에서 가장 우수한 성능을 보인 주위기반의 딥러닝 모델은 기계가 SNS의 문장 또는 영상을 다른 양식으로부터 스스로 생성할 수 있는데 활용될 수 있다.

Key Words : Cross Modal Retrieval, Social Network Service(SNS), Text-Image Dataset, Deep Learning, Computer Vision

ABSTRACT

Recently, various types of large dataset give the great help for the deveopment of AI technology as the same way as ImageNet does for the computer vision area. Unfortunately, however, there is no such large dataset for cross-modal retrieval between the text of Korean language and color images. This paper proposes the method how to easily collect and arrange to construct such dataset for cross-modal retrieval from Instagram, a kind of SNS(Social Network Service). Then we construct the dataset according to the methos, and the applicability has been proven by performing the cross-modal retrieval experiments. In the experiment, several methods for cross-modal retrievals are adopted including attention-based deep learning approach to compare the performances. The dataset in the study can be used to various multi-modal machine learning which requires the analysis of short Korean sentences, and the attention-based deep learning model which provides the best performance can be applied to automatically generate a Korean sentence from a color image, or a color image from a Korean sentence in SNS.

※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2015R1D1A1A01058062)

◆ First Author : Department of Computer Engineering, Chonbuk National University, late1704@naver.com, 학생회원
° Corresponding Author : Department of Computer Engineering, Chonbuk National University, chlee@chonbuk.ac.kr, 종신회원

논문번호 : 201902-458-C-RN, Received February 12, 2019; Revised May 10, 2019; Accepted May 13, 2019

Ⅰ. 서 론

ImageNet[1]과 같은 데이터셋은 컴퓨터 비전 기술의 비약적 발전에 있어 주요 원동력이 되어 왔다. 이러한 대규모의 데이터셋은 수백만 개의 매개변수를 가진 딥러닝 모델이 적절히 학습될 수 있도록 지원하여, 데이터에 대한 분석 수준을 인간과 비슷한 수준으로 향상시키는 부분에서 중요한 역할을 수행하고 있다. 대규모 데이터셋이 존재하는 연구 분야 중 텍스트 분석의 경우는 GoogleNews 데이터셋이 도움을 제공하고 있다. 하지만 해당 데이터셋은 영문 텍스트와 관련되어 있기 때문에 현재 한국어를 위해 공개된 대규모 데이터셋은 존재하지 않은 실정이고, 이에 따라 문장과 칼라영상 간 교차양식 검색을 위한 한국어 문장 데이터셋 역시 구축되어 있지 않다.

본 연구의 경우 이러한 상황을 타개하는 데 SNS 게시물의 이용을 고려해 보았다. 최근 스마트폰 보급의 증가로 인해 누구나 SNS를 쉽게 접할 수 있게 되었고, 자연스럽게 SNS에 올라오는 게시물의 수 또한 기하급수적으로 증가하게 되었다. 따라서 본 논문은 현재의 상황을 활용하는 방안으로써, SNS의 일종인 인스타그램(Instagram) 게시물을 통해 대규모의 데이터셋을 손쉽게 구축할 수 있는 방법을 제안하였다.

인스타그램 게시물은 이미지, 비디오, 텍스트, 오디오와 같은 다중양식(multi-modal) 데이터들을 포함하기 때문에, 이들 게시물로부터 다중양식 데이터를 손쉽게 수집할 수 있다. 본 연구에서는 이 중에서 텍스트와 영상의 쌍으로 이루어진 데이터를 수집함으로써 텍스트-영상 다중양식 기계학습 연구에 적합한 데이터셋을 구축하였다.

인스타그램으로 부터의 데이터 수집과정은 인스타그램 게시물을 검색하기 위한 해시태그를 이용한다. 제안한 방법은 그 대상 및 범위에 제약이 따르지 않으나, 편의상 논문에서는 그 범위를 제한하여 자연영상을 대상으로 하는 해시태그를 활용하였으며, 불필요한 데이터를 제거하고 총 272,400개의 영상-텍스트 쌍을 수집하였다.

한편 최근 비약적으로 발전하고 있는 다중양식 기계학습의 한 분야인 교차양식 검색(cross-modal retrieval)은 다양한 교차양식 데이터 중에서 하나의 양식을 통해 다른 양식에 대한 검색을 수행한다. 본 연구의 경우, 텍스트-칼라영상을 수집하여 구축한 다중양식 인스타그램 데이터셋의 활용성을 확인하기 위한 방안으로 교차양식 검색을 수행해 보았다.

일반적으로 교차양식 검색은 각각의 양식을 연계하는 방법에 따라 다양한 방식으로 수행될 수 있다. 본 연구에서는 주의(attention)기반 DNN(Deep Neural Network)방법을 포함하여 다양한 방법으로 텍스트와 칼라영상을 연계시키고 그 성능을 평가해 보았다. 그림 1은 본 연구에서 활용된 주의기반 텍스트-칼라영상의 교차양식 검색의 예를 보여주고 있다.

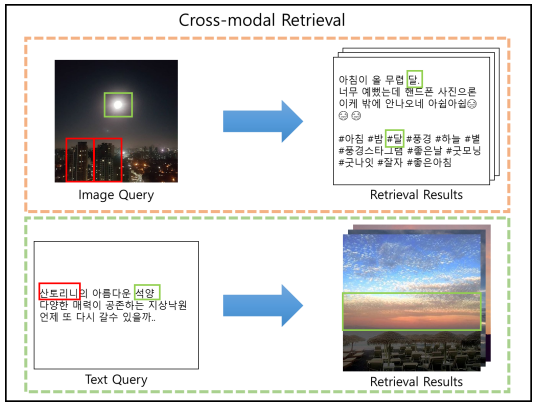

그림 1. 교차 양식 검색의 예
Fig. 1. An example of cross-modal retrieval

텍스트-칼라영상 쌍으로부터 교차양식 검색을 가능하도록 하려면 각 양식의 벡터 특징추출과 추출된 서로 다른 양식들에 대한 특징을 연계시키는 과정이 필요하다. 본 실험에서 영상 데이터에 대해서는 VGG-19[2]를 이용하여 벡터 특징을 추출하였다. 또한 텍스트의 경우 우선적으로 형태소 분석을 수행하고, 학습된 Word2Vec[3] 모델로 벡터 임베딩을 수행한 뒤 WordCNN[4]을 통해 학습된 특징을 추출하였다.

본 논문에서는 이렇게 추출된 두 양식의 특징을 가지고 3가지 전통적인 교차양식 연계방법과 5가지 DNN기반의 연계방법을 이용, 교차양식 검색을 수행하고 성능을 비교해 보았다. 또한 성능비교를 통해 우수한 성능을 보였던 MCSM (Modality-specific cross-modal similarity measurement)[5] 알고리즘에 대해 여러 가지 세부실험과 알고리즘에 대한 검증을 수행하였다. 성능이 검증된 MCSM 모델 등은 SNS에서 기계가 SNS의 문장 또는 영상을 다른 양식으로부터 스스로 생성하는 작업에 활용될 수 있다.

이 논문의 2장에서는 기존 데이터셋과 문장-칼라영상 쌍을 가지고 교차양식 검색을 수행하는 기존 연구들의 검토하였고, 3장에서는 데이터셋의 구성방법을 설명하였다. 이후 4장에서는 대규모 데이터셋를 활용한 교차양식 검색 딥러닝 모델의 학습과 실험결과를 보였으며, 마지막 5장에서는 결론 및 향후 계획을 설명한다.

Ⅱ. 관련 연구

본 절에서는 기계학습과 관련된 기존 대규모 데이터셋들을 검토하고, 공통 벡터공간 학습을 사용하는 교차양식 검색 방법 및 DNN에서의 주의 체제와 관련된 최근 연구동향을 살펴보았다.

2.1 기존 데이터셋 및 한국어 문장-영상 데이터셋

영상 데이터셋들은 영상에 대한 이해 면에서 컴퓨터 비전 알고리즘의 발전에 크게 기여하고 있다. 컴퓨터 비전 연구는 Caltech 101/256 [6,7], MSRC[8], PASCAL[9]과 같은 소규모 데이터셋으로부터 시작하여 ImageNet과 같이 보다 큰 데이터셋을 활용하는 방향으로 발전하고 있다. 일반적으로 커다란 규모의 딥러닝 모델은 방대한 매개변수를 적절히 학습시키는데 어려움이 따른다. 그러나 AlexNet[10] 이나 Inception [11] 은 큰 규모를 가진 모델임에도 불구하고 수백만 개의 주석이 달린 ImageNet 데이터셋을 활용함으로써 성공적으로 학습이 될 수 있었다.

텍스트 분야에서는 Google News처럼 대규모 데이터 셋을 기반으로 학습된 Word2Vec를 공개적으로 쉽게 얻을 수 있지만, 한국어 기반으로 하는 공개적인 데이터셋은 존재하지 않는다. 또한 이미지-텍스트 문서 쌍으로 영어로 된 데이터로는 Wikipedia dataset [12], Pascal Sentence dataset[13]가 있지만, 이 부분 역시 한국어로 된 이미지-텍스트 문서 쌍은 찾아보기 어렵다.

2.2 교차양식 검색을 위한 공통 벡터공간 학습

교차양식 검색을 위해서는 서로 다른 두 양식을 연계시키는 과정이 필요하다. 최근 사용되는 교차양식 검색 방법은 서로 다른 양식의 특징이 공통 벡터공간에서 매핑될 수 있도록 유도하고, 교차양식의 유사도를 그 공통 공간에서 측정함으로써 두 양식간의 연계를 구현한다. 이러한 교차양식의 연계는 전통적인 통계 상관 분석 방법[14], 교차 양식 그래프 정규화 방법, DNN 기반 방법 등이 활용되고 있다.

2.2.1 전통적인 통계적 상관 분석 방법

전통적인 통계적 상관분석 방법은 공통 공간학습 방법으로서 서로 다른 양식의 특징을 공통 공간에 투영하여 벡터표현을 얻었다. 즉 통계적 상관분석 방법에서는 통계적 정보를 최적화하는 방식으로 공통 공간 벡터를 얻을 수 있는 선형 투영행렬(linear projection matrices)을 학습하게 된다.

대표적인 통계적 상관 분석 방법으로는 정준 상관분석 (Canonical Correlation Analysis, CCA)[15]이 있으며 문장-칼라영상의 쌍과 같은 서로 다른 데이터 양식간의 쌍 상관(pairwise correlation)을 최대화하는 투영행렬을 얻는 것이 목표이다. 또한, 교차양식 검색의 정확성을 향상시키기 위해 CCA를 확장하여 의미정보를 통합하기도 한다. 예를 들어 Costa et al.[16] 의미 범주 레이블을 통합하여 CCA의 성능을 향상시켰으며, Multi-view CCA[17]는 높은 수준의 의미를 영상과 텍스트 정보에 이어 세 번째 관점으로 포함하여 성능 향상을 도모하였다. Ranjan et al.[18]의 다중 레이블 CCA는 다중 레이블 주석의 형태로 상위 수준의 의미 정보를 함께 고려한다. 또 다른 방법 중 하나인 교차양식 인자분석(Cross Modal Factor Analysis, CFA)[19]은 서로 다른 데이터의 양식을 하나의 공통 공간에 투영 한 후 프로베니우스 노름(Frobenius Norm) 최소화하는 것을 목적함수로 한다.

2.2.2 교차양식 그래프 정규화 방법

그래프 정규화[20]는 준 감독 학습(semi-supervised learning)에서 레이블된 그래프를 구성하는 데 사용된다. Zhai et al.[21]의 연구는 그래프 정규화를 교차양식 검색에 통합한 최초의 시도이며, JGRHML(Joint Graph Regularized Heterogeneous Metric Learning) 라 이름붙인 메트릭 학습을 통해 서로 다른 양식의 구조를 하나의 합동 그래프 정규화에 통합한다. 또한, JRL (Joint Representation Learning)[22]에서는 서로 다른 양식에 대한 여러 그래프를 구성하고 상관관계 정보와 의미정보를 함께 고려하여 투영 행렬을 학습하는 방법을 제안한다. 뿐만 아니라 Peng et al.[23]는 다섯 개의 양식에 대한 공통 공간을 학습하기 위해 통합 된 하이퍼 그래프를 구성하여 이전의 연구[20]의 성능을 더욱 향상 시키는 동시에 세분화된 정보를 활용할 수 있게 하였다.

2.2.3 DNN 기반 방법

DNN은 비디오 분류[24], 객체인식[25]과 같은 컴퓨터 비전 문제 해결에 엄청난 기여를 하고 있다. 특별히 교차양식 검색에서는 공통적인 벡터공간으로의 매핑을 위해 DNN이 가지는 비선형성을 활용한다. 대부분의 DNN 기반 방법은 영상과 텍스트 같은 다양한 양식에 대해 각각의 하위 네트워크를 구성하여 각각의 특징을 추출하고 이들 특징의 교차양식 상관관계를 모델링하기 위한 공통 벡터공간으로 결합 계층(joint layer)에서 연결된다. Bimodal Autoencoders

(Bimodal AE)[26]는 Restricted Boltzmann Machine (RBM)을 확장하여 재구성 오류(reconstruction error)를 최소화함으로써 다중 양식을 모델링하기 위해 제안되었다.

Srivastava et al.[27]은 Multimodal Deep Belief Network (Multimodal DBN)을 제안했는데, 여기서는 두 가지 유형의 DBN을 사용해서 기존의 특징에 대한 분포를 모델링하고, 결합 분포 모델링과 공통 표현을 얻기 위해서 결합 RBM을 사용했다.

Correspondence Autoencoder (Corr-AE)[28]와 Deep Canonical Correlation Analysis (DCCA)[29]는 두 개의 서브 네트워크로 구성된다. Corr-AE는 상관관계 정보와 재구성(reconstruction) 정보를 모델링하고, DCCA는 DNN과 CCA를 결합하여 두 개의 서브 네트워크 상단에서 상관관계를 최대화한다.

Peng et al.[30]은 2단계 학습 프레임워크에서 양식 사이의 상관관계와 양식 내의 상관관계를 모델링하기 위해 CMDN(Cross-media Multiple Deep Network)을 제안하였다. 이들 연구들에서는 주로 직접 설계한 특징(hand-crafted features)을 이미지 입력으로 사용한다. 반면에 Wei et al.[31]은 CNN을 통해 각 양식의 특징을 추출하고 이를 높은 단계의 의미매칭(semantic matching)을 통해 연계하는 심층적인 의미매칭 Deep-SM을 제안하였다. 이는 CNN의 강력한 표현 능력을 이용하여 검색 정확도를 향상시킨다. He et al.[32]는 딥러닝과 양방향 표현 학습을 통해 일치하는 영상-텍스트 쌍과 일치하지 않는 영상-텍스트 쌍을 모델링하기 위해 두 개의 CNN 기반 네트워크를 채택한다.

2.3 주의 체제 (Attention Mechanism)

최근 DNN의 발전방향 중 하나인 주의 체제는 모델이 시각적 입력과 텍스트 입력의 세부 부분에 초점을 맞출 수 있게 해주며, 다중양식(multimodal) 과제에 적용되고 있다. 본 절에서는 주의 체제 중 본 연구와 관련된 시각 주의(visual attention)와 텍스트 주의(textual attention)를 검토한다.

2.3.1 시각 주의

최근 영상의 특정 부분에 많은 관심을 기울인 상태에서 여러 영상처리 작업을 수행하는 시각적 주의 모델을 활용한 많은 방법들이 소개되고 있다. Mnih et al.[33]은 순환 신경망 (Recurrent Neural Network, RNN)을 채택하여 이미지 분류를 위해서 주의 기반의 시각 처리 프레임워크를 제안했다. 또한 Gregor et al.

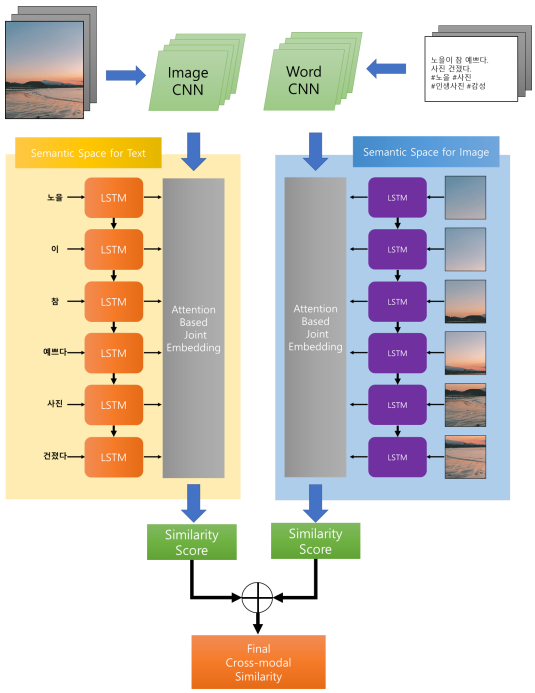

그림 2. MCSM 알고리즘의 간단한 도식화
Fig. 2. Block diagram of MCSM algorithm

[34]는 영상생성을 수행하기 위해 자동 인코딩 (Auto-encoding) 프레임워크를 설계하는 데에 공간 주의 체제를 제안한다. Yang et al.[35]는 영상 질의응답을 위한 스택 주의 네트워크 (Stacked Attention Networks, SAN)를 제안한바 있는데 이 모델은 쌓여진 주의 모델을 통해서 관련 영상 영역을 질문과 일치시킬 수 있다.

2.3.2 텍스트 주의

자연어 처리 분야의 일부 연구들은 인코더-디코더 네트워크의 의미론적 정렬(semantic alignments)을 찾기 위해 텍스트 주의 모델을 활용했다. Rocktaschel et al.[36]은 쌍으로 구성된 단어나 어구의 수반관계 (entailments)을 추론하는 단어 단위 신경망 주의 체제를 제안한다.

한편 Hermann et al.[37] 복잡한 질문을 읽고 답변하는 것을 배우는 주의 기반 심층 신경망을 개발했다. Rush et al.[38]은 입력된 문장에 따라 요약을 생성하기 위해 지역 주의기반 모델을 채택한 데이터 접근 방식을 제안한다.

MCSM[5]에서는 주의 체제를 통해 양식 특징을 찾아낼 수 있고, 서로 다른 양식에 대한 독립적인 의미 공간을 구성한다. 즉 각 양식 전체의 공통된 벡터 표

현 없이 종단 간 프레임 워크(end-to-end framework)를 통해 주의 체제를 활용 의미 공간에서의 교차양식 유사도를 직접 생성한다. 그림 2는 모델개관을 도식적으로 보여준다.

Ⅲ. 데이터셋 구성 및 특징추출

본 절에서는 대규모의 텍스트-칼라영상 데이터셋을 구성하는 절차와 구성된 데이터셋에서 텍스트 부분과 칼라영상부분의 특징을 추출하는 방법을 서술한다.

3.1 인스타그램 게시물 수집

본 연구에서는 인스타그램 게시물로부터 데이터를 수집하고 텍스트-칼라영상 데이터셋을 구축하였다.

인스타그램은 편리하게 데이터를 수집할 수 있는 데이터 창고이며, 게시물은 생활 속 다양한 주제를 담고 있다. 본 연구에서는 그 주제들 중에 주로 바다, 산, 하늘 등의 자연물을 주제로 텍스트-영상 데이터를 수집하였다. 해당 주제는 하나의 예에 불과하며 다른 주제를 선정한다면 다른 종류의 데이터를 수집할 수 있다.

일반적으로 인스타그램의 텍스트는 사용자의 주관적인 견해를 반영하고 있으며, 사용자는 해시태그 시스템을 이용하여 게시물과 주제를 연결시킬 수 있다. 또한, '#바다,#산,#하늘' 등의 형태로 부가된 해시태그는 메타데이터로 활용할 수 있으며, 이렇게 작성된 메타데이터는 인스타그램의 검색 시스템을 통하여 게시물을 검색하고, 이와 관련된 여러 가지 관련 콘텐츠를 얻는데 활용 된다. 즉 해시태그를 통해 주요 주제를 간결하게 나타내고 검색에 클래스로 이용할 수 있게 하는 것이다.

표 1은 본 연구를 통해 수집한 주제들의 상위와 하위클래스들을 보여준다. 즉 본 연구에서는 이들 25개의 주제들에 해당하는 해시태그를 이용하여 게시물을 수집하고 이를 필터링하여 총 272,400 영상-텍스트 쌍을 수집하였다.

표 1. 상위 클래스에 따른 하위 클래스와의 관계
Table 1. Relationship to subclasses by superclass

Ancestor class	Subclass				
바다	해변	해안	강		
하늘	구름	먹구름			
밤하늘	달	별			
산	정상	호수	나무	언덕	초원
노을	일출	일몰			
자연	풍경	경치			
여행	관광	휴가			

3.2 수집 데이터의 필터링

인스타그램의 검색 시스템을 통하여 검색된 게시물은 교차양식 검색에 활용하기 부적합한 데이터도 포함되어 있다. 따라서 다음과 같은 방법으로 데이터를 필터링하여 목적에 맞는 데이터를 선별하였다. 만약 교차양식 검색이 아닌 다른 목적으로 데이터를 수집한다면 이러한 필터링은 다른 방법으로 진행할 수도 있다.

먼저 영상 크기가 224x224의 VGG-19 입력보다 작은 영상 게시물과 동영상-텍스트 쌍을 제외하였다. 즉 영상의 크기가 너무 작아 교차양식 검색을 위한 특징추출이 불가능하거나 동영상은 본 연구의 대상이 아니기 때문에 제거하였다. 또한 본 연구의 데이터셋은 교차양식 검색용이기 때문에 사용자의 게시물에서 영상 또는 텍스트 중 하나라도 없는 게시물은 모두 제외하였다. 즉 영상의 크기가 CNN 특징추출기인 VGG-19의 입력크기 보다는 작지 않고, 텍스트와 영상 모두를 담고 있는 게시물만을 추출하였다.

한편 해시태그를 이용하여 검색할 때 광고 게시물 등이 존재할 수 있다. 또한, 표 1의 범위를 축소한 연구대상의 클래스와 맞지 않는 게시물은 제외하였다. 즉 본 연구의 데이터 수집 과정에서 광고 계정은 따로 블랙리스트 만들어 해당 게시물들은 수집할 때 제외하게 만들었으며, 기타 관련이 없는 게시물은 영상과 텍스트를 검토하여 제외하였다.

3.3 데이터셋의 구성

이렇게 수집 정제된 인스타그램 데이터는 게시물 검색에 사용한 표 1의 해시태그를 기준으로 25개의 하위 클래스와 7개의 상위클래스로 구분하였다. 여기서 하위 클래스는 상위클래스에 속하거나, 유사한 용어이거나 밀접한 관련을 가진 용어들로 정의하였다. 본 연구에서는 총 272,400개의 영상-텍스트 쌍을 최종 수집된 데이터셋으로 하였으며 이들은 각각 7개의 상위클래스와 8개의 하위클래스에 속하게 분류된다.

데이터셋의 영상-텍스트 쌍에서 텍스트는 여러 개의 단어로 이루어진 문장 또는 문장들로 구성되며, 영상은 한 장의 영상을 포함한다. 그림 3은 수집된 272,400개의 영상-텍스트 쌍의 하위클래스에 따른 데이터 수를 표현한다.

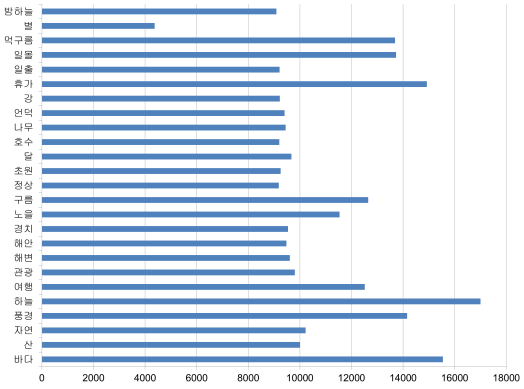

그림 3. 각 클래스별 수집된 게시물의 수
Fig. 3. Number of posts collected for each class

3.4 특징 추출

본 연구에서는 구축된 데이터셋의 영상 및 텍스트에 대하여 각 DNN을 이용한 학습된 특징을 추출하고 이를 이용하여 교차양식 검색을 수행한다. 텍스트의 특징은 형태소 분석기를 거친 뒤, 학습된 Word2Vec를 이용하여 WordCNN을 통해 특징을 추출한다.

3.4.1. 칼라영상의 학습된 특징 추출

영상의 학습된 특징은 ImageNet 데이터로 사전 학습된 VGG-19를 이용하여 학습된 특징을 추출하였다. 구체적으로, 각 이미지를 256x256의 크기로 전처리를 하고, VGG-19의 마지막 풀링 계층(pooling layer5)에서 특징을 추출하였다. 마지막 풀링 계층은 512개의 필터를 통해서 해당 이미지의 512차원의 특징 벡터를 얻을 수 있다.

3.4.2 텍스트를 위한 특징 추출

수집된 텍스트는 대부분 한글로 이루어져 있다. 영어의 경우에는 띄어쓰기를 기반으로 자른 후 사용해도 무관하지만, 한글의 경우에는 복잡한 언어 구조상의 이유로 형태소 분석이 필수적이기 때문에 한국어 정보처리를 할 수 있는 패키지인 KoNLPy(Korean NLP)를 이용하였다.

KoNLPy가 제공하는 형태소 분석기는 꼬꼬마, 코모란, 트위터 세 가지 분석기가 있는데, 이들의 성능 분석은 4장의 실험 및 분석 부분에서 언급한다. 형태소 분석이 끝난 뒤 단어의 벡터 임베딩을 위해서 Word2Vec를 이용하였다. Word2Vec의 학습 데이터로 약 210만개의 문장으로 이루어진 한국어 위키피디아 데이터, 약 110만개의 문장으로 이루어진 네이버 영화 리뷰 데이터, 인스타그램으로부터 구성된 약 272,400개의 텍스트 데이터를 합친 데이터셋을 이용하여 학습을 통해 구성하였다.

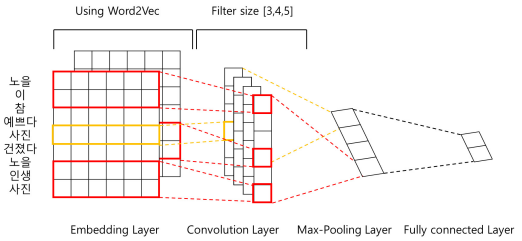

그림 4. 텍스트 특징 추출을 위한 WordCNN 구조
Fig. 4. WordCNN structure for feature extraction

구성된 영상-텍스트의 데이터셋에서 텍스트, 즉 문장들은 사용된 단어들을 Word2Vec로 벡터임베딩하고 WordCNN을 통해 텍스트 특징을 추출한다. 본 연구에서는 Word2Vec[3] 모델을 이용하였으며, 사용된 WordCNN[4]의 구조는 그림 4와 같다.

Ⅳ. 실험 및 분석

이 장에서는 구축된 텍스트-영상 교차양식 데이터셋의 활용가능성을 검토하고자 교차양식 검색 실험을 수행하였다. 교차양식 검색에는 여러 알고리즘을 적용하여 그 성능을 평가하였으며, 평가지표로는 Mean Average Precision(mAP)와 Recall@K를 사용하였다. 그리고 가장 우수한 성능을 보인 MCSM 알고리즘에 대해서 여러 가지 실험을 추가적으로 수행하였다. 데이터셋은 학습 집합과 테스트 집합으로 나누어 그 구성비가 약 7:3을 이루도록 하였다. 또한 테스트 데이터 중 일부는 Recall@K를 계산하기 위해 사용된다.

4.1 비교된 알고리즘

CCA[15], CFA[19], KCCA(Kernel CCA)[39] 같은 3가지 전통적인 교차양식 검색 방법과 DCCA[17], CMDN[30], Deep-SM[31], CMPL[40], MCSM[5]과 같은 5가지 DNN 기반의 방법을 통해 실험을 진행하였다. 해당 방법들에 대해 간략한 소개는 다음과 같다.

- CCA[15]는 하나의 공통 벡터공간에서 다른 양식의 투영된 특징 간의 상관관계를 최대화하기 위해 행렬을 학습한다.
- CFA[19]는 하나의 공통 벡터공간으로 투영 한 후 서로 다른 양식의 데이터 간에 프로베니우스 노름(Frobenius Norm)을 최소화한다.
- KCCA[39]는 커널 기능을 사용하여 특징을 고차원

공간으로 투영 한 다음 CCA로 공통 벡터공간을 학습한다. 실험에서는 가우시안 커널을 커널 함수로 채택했다.

- DCCA[17]는 CCA의 비선형 확장이다. 상관관계는 두 개의 개별 서브 네트워크의 출력 계층 간에 최대화된다.
- CMDN[30]은 여러 개의 깊은(deep) 네트워크를 사용하여 별도의 표현을 생성하고 누적 된 네트워크로 공통 표현을 학습한다.
- Deep-SM[31]은 의미론적 매칭을 수행하기 위해 영상의 경우 CNN 전달학습(transfer learning)을 채택하며 특징을 추출하며, 텍스트의 경우는 직접 설계된 특징을 FCN(fully connected network)로 추상화하여 활용한다.
- CMPL[40]는 특징벡터를 얻기위해 텍스트에 대해서는 Bi-directional LSTM을 수행하고, 이미지에 대해서는 CNN을 수행한 뒤 결합 임베딩을 학습한다.
- MCSM[5]은 주의 체제를 통해 양식의 특징을 찾아내고, 서로 다른 양식에 대한 독립적인 의미 공간을 구성한다.

4.2 교차양식 검색 알고리즘의 비교

본 연구의 실험은 다음과 같이 정의 된 두 종류의 검색 작업으로 구성된다.

- 이미지 쿼리 텍스트 (I2T: 이미지 → 텍스트) : 이미지 쿼리를 사용하여 계산 된 교차양식 유사도로 순위가 매겨진 평가 집합에서 관련 텍스트를 검색한다.
- 텍스트 쿼리 이미지 (T2I: 텍스트 → 이미지) : 텍스트 쿼리를 사용하여 계산 된 교차양식 유사도로 순위가 매겨진 평가 집합에서 관련 이미지를 검색한다.

4.2.1 성능평가 지표

또한 본 연구에서는 다음 두 평가지표를 이용하여 성능을 평가하였다.

- Mean Average Precision (mAP) : Average Precision은 다음과 같이 정의된다.

$$AP = \frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k} \times rel_k \quad (1)$$

여기서 검색결과는 개의 요소를 포함한다. 는 상위 번째 검색결과와 관련된 요소의 수이다. 는 번째 검색결과가 질의와 관련이 있을 때 1로 설정되고, 그렇지 않으면 는 0으로 설정된다. 또한 은 전체 검색 중 관련된 요소의 수를 의미한다.

mAP는 전체 질의 수에 대한 AP의 평균값으로 검색 결과의 순위와 정밀도를 동시에 고려하며, 교차양식 검색 작업에서 광범위하게 채택된다. MAP 점수는 [28]에서와 같이 검색에서 채택된 상위 50개뿐 아니라 모든 반환 된 결과에 대해 계산된다.

- Recall@K : 교차 양식 검색을 평가를 위해 많은 연구에서 Recall@K (K = 1, 5, 10) [41]을 채택한다. Recall@K 는 상위 K개의 검색 결과 중에서 적어도 하나의 ground-truth가 검색되는 쿼리의 비율을 나타낸다.

인스타그램 데이터셋에서는 클래스 별로 1000개의 영상-텍스트 쌍을 ground-truth로 두고 실험을 진행하였다.

4.2.2 교차양식 검색알고리즘 성능비교

표 2는 8개의 알고리즘의 mAP 값을 보여주며 MCSM 알고리즘이 8개의 알고리즘에서 가장 좋은 검색 정확도를 나타내는 것을 알 수 있다. 이는 동일한 MCSM 알고리즘을 Wikipedia 데이터셋에 적용한 결과에 비해 약 7.5% 높게 나타났는데 이는 본 연구의 경우 대상을 표 1과 같이 제한하였기 때문으로 해석된다. 또한, 표 3도 8개의 알고리즘의 Recall@K (1, 5, 10)의 결과 값을 보여주며 모든 부분에서 MCSM 알고리즘이 우수함을 보였다. 특별히 표 2와 3의 결과들은 본 연구에서 수집한 데이터셋이 학습시에 과적합(overfitting) 등과 같은 문제를 발생하지 않을 정도로 충분한 양이며, 비록 자연이라는 제한적인 범위의

표 2. 알고리즘에 대한 mAP 결과
Table 2. MAP results for the algorithm

Method	mAP scores		
	I2T	T2I	Avg
MCSM	**0.579**	**0.551**	**0.565**
CMPL	0.517	0.495	0.488
CMDN	0.524	0.506	0.515
Deep-SM	0.539	0.519	0.529
DCCA	0.549	0.489	0.519
KCCA	0.461	0.457	0.459
CFA	0.453	0.449	0.451
CCA	0.183	0.188	0.186

표 3. 알고리즘에 대한 Recall@K 결과
Table 3. Recall @ K results for the algorithm

Method	Recall@1			Recall@5			Recall@10		
	I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
MCSM	**0.581**	**0.515**	**0.548**	**0.796**	**0.627**	**0.712**	**0.897**	**0.779**	**0.838**
CMPL	0.558	0.477	0.518	0.734	0.609	0.672	0.876	0.766	0.821
CMDN	0.563	0.492	0.528	0.757	0.616	0.687	0.852	0.755	0.804
Deep-SM	0.577	0.501	0.539	0.764	0.620	0.692	0.863	0.761	0.812
DCCA	0.570	0.499	0.535	0.761	0.618	0.690	0.861	0.757	0.809
KCCA	0.501	0.452	0.477	0.711	0.581	0.646	0.801	0.718	0.760
CFA	0.495	0.397	0.446	0.681	0.541	0.611	0.756	0.683	0.720
CCA	0.365	0.247	0.306	0.622	0.534	0.578	0.733	0.668	0.701

데이터라 할지라도 교차양식 검색에 성공적으로 활용할 수 있음을 보여주고 있다.

4.3 MCSM 알고리즘에 대한 보충실험

앞선 교차양식 검색실험에서 가장 성능이 좋은 알고리즘은 MCSM 알고리즘이었다. 이에 MCSM의 교차 양식 검색 결과를 출력하고, 형태소 분석기에 따른 실험을 진행하였다. 또한, Word2Vec의 차원을 변경하거나 Word2Vec를 사용하지 않는 추가적인 실험을 진행하고, Word2Vec가 잘 학습 되었는지 t-SNE를 통해 가시화를 수행하였다. 마지막으로 알고리즘의 중간 과정의 주의 망(attention network) 부분에서 네트워크가 어떻게 동작했는지 가시화를 수행하였다.

4.3.1 MCSM 알고리즘의 교차 양식 검색 결과

그림 5는 인스타그램 데이터를 통한 교차 양식 검색의 결과들 중 유사도가 높은 10개의 항목을 출력한 결과이다. 먼저 이미지 “바다”를 넣었을 때 검색 결과들은 해시태그 “#바다”를 공통적으로 가지고 있었으며 종종 등장하는 단어로서 제주도와 관련된 단어들의 빈도가 높았다.

이미지 “구름”을 넣었을 때 바다와 마찬가지로 해시태그 “#구름”을 공통적으로 가지고 있었으며 연관성이 높은 단어로는 “날씨”, “예쁘다”, “하늘”이라는 단어가 종종 나타났다. 다음으로 “노을”과 관련된 텍스트를 넣었을 때 대부분 붉게 물든 하늘인 노을이 검색 결과로 나왔고 노란빛을 띄는 노을도 나타났다. 그리고 “산”과 관련된 텍스트를 넣었을 때 대부분 산의 이미지가 나타났고, 사람과 함께 찍은 산 또한 검색 결과로 나타났다.

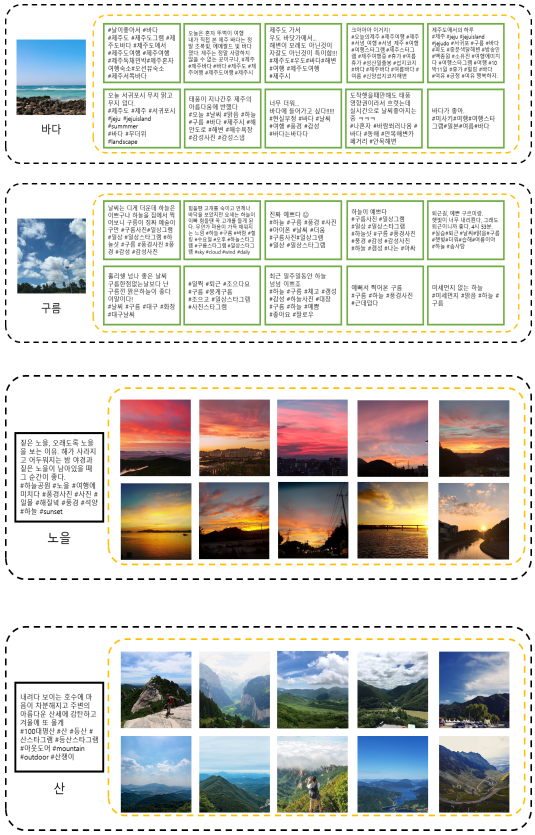

그림 5. MCSM 알고리즘의 교차 양식 검색의 결과
Fig. 5. Result of cross-modal retrieval of MCSM algorithm

4.3.2 형태소 분석기에 따른 결과

한국어 정보처리를 할 수 있는 패키지인 KoNLPy(Korean NLP)[42]에서 제공하는 형태소 분석기 종류에는 꼬꼬마, 코모란, 트위터가 존재하며 속도는 트위터, 코모란, 꼬꼬마 순으로 빠르다. 표 4에서는 3가지 형태소 분석기에 따른 Precision이고, 표 5는 Recall@K를 구한 결과이다. 3가지 분석기의 Precision, Recall@K 모두 비슷한 결과를 가진다. 따라서 속도가 가장 빠른 트위터 분석기가 효율적임을 알 수 있다.

표 4. 형태소 분석기에 따른 Precision 결과
Table 4. Precision Results by Morpheme Analyzer

Analyzer	Precision		
	I2T	T2I	Avg
꼬꼬마	**0.581**	0.550	**0.566**
코모란	0.577	0.547	0.562
트위터	0.579	**0.551**	0.565

표 5. 형태소 분석기에 따른 Recall@K 결과
Table 5. Recall@K Results by Morpheme Analyzer

Analyzer	Recall@1			Recall@5			Recall@10		
	I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
꼬꼬마	0.580	0.513	0.547	**0.798**	0.626	**0.712**	**0.898**	**0.781**	**0.840**
코모란	0.579	0.514	0.547	0.794	0.628	0.711	0.874	0.780	0.827
트위터	**0.581**	**0.515**	**0.548**	0.796	**0.627**	**0.712**	0.897	0.779	0.838

4.3.3 Word2Vec의 차원에 따른 결과

Word2Vec는 각각 100, 300, 500차원을 사용하여 Skip-gram[3]을 학습했으며, 학습 결과의 검증을 위해 5-폴드 교차검증(5-fold cross validation)을 사용하였다. 또한, 추가적으로 Word2Vec를 사용하지 않고 랜덤으로 준 값으로 단어를 임베딩 하여 실험을 진행 하였다. 표 6는 Word2Vec에 따른 Precision값이 정리하였다. 4가지 실험 중 I2T (image to text)에서는 300차원일 때 가장 좋은 결과가 나왔으나 T2I (text to image)와 평균값에서는 500차원일 때 가장 좋은 결과가 나왔다.

또한 표 7에서는 Recall@5의 T2I와 Recall@10의 T2I의 결과를 제외한 모든 부분에서 500차원 보다 300차원의 값이 좋은 결과로 나타났다. 두 가지 Precision과 Recall@K를 비교했을 때 300차원이 가장 좋은 결과를 나타냈으며, 500차원이 더 좋은 결과를 나타냈을 때 큰 차이가 나지 않았다. 그래서 계산량을 고려할 때, 300차원이 더욱 효율적임을 알 수 있다.

표 6. Word2Vec에 따른 Precision
Table 6. Precision according to Word2Vec

dimension	Precision		
	I2T	**T2I**	**Avg**
d-100	0.552	0.537	0.545
d-300	**0.579**	0.551	0.565
d-500	0.575	**0.562**	**0.569**
Random	0.498	0.513	0.506

표 7. Word2Vec에 따른 Recall@K
Table 7. Recall@K according to Word2Vec

dimension	Recall@1			Recall@5			Recall@10		
	I2T	**T2I**	**Avg**	**I2T**	**T2I**	**Avg**	**I2T**	**T2I**	**Avg**
d-100	0.564	0.493	0.529	0.751	0.610	0.681	0.855	0.731	0.793
d-300	**0.581**	**0.515**	**0.548**	**0.796**	0.627	**0.712**	**0.897**	0.779	**0.838**
d-500	0.579	0.513	0.546	0.781	**0.633**	0.707	0.873	**0.781**	0.827
Random	0.497	0.453	0.475	0.692	0.557	0.625	0.756	0.684	0.72

4.3.4 t-SNE를 통한 학습된 Word2Vec의 검증

본 실험에서는 Word2Vec의 성능을 확인하기 위해 t-SNE를 통하여 300차원으로 학습된 Word2Vec를 시각화해 보았다. 임베딩 초기화 방법으로는 PCA를 사용하였으며, 빈도수가 높은 상위 50개의 명사로 t-SNE를 한 결과는 그림 6와 같다. 우선 왼쪽 상단 부분에는 "하늘"과 "구름" 주변으로 관련 단어들이 존재를 하고 있으며 아래쪽에는 "바다"와 관련된 단어들이 분포한다.

"바다"와 "하늘" 사이에는 "푸른"과 "파란"을 가운데에 두고 있다. 다음으로 "하늘"의 오른쪽에는 "밤하늘", "야경"과 같은 단어들이 모여 있으며, 그 위에는 "석양", "노을"과 관련된 단어들이 모여 있다. 마지막으로 오른쪽 아래에는 "산"과 관련된 단어들이 모여 있다. 따라서 이 가시화 실험의 결과로부터 학습된 한국어 Word2Vec 표현은 의미가 잘 보존되고 있음을 확인 가능하다.

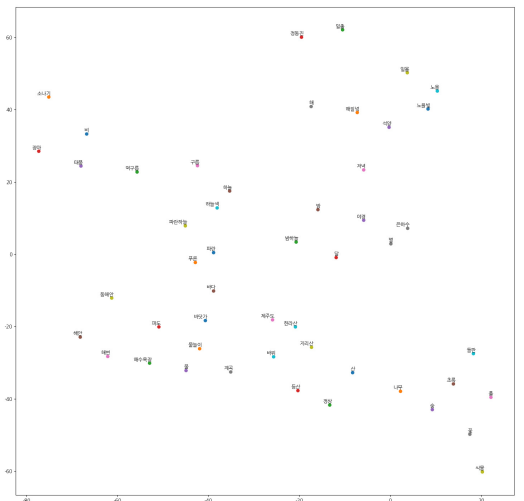

그림 6. t-SNE를 이용한 Word2Vec 가시화
Fig. 6. Visualization of Word2Vec using t-SNE

4.3.5 주의 망를 통한 결과

MCSM 알고리즘에서는 영상부와 텍스트부의 LSTM결과 값을 입력으로 Softmax를 통해 결과가 나온 값을 활용하여 영상과 텍스트에 대해서 해당 모델이 어느 부분에서 높은 결과 값이 나왔는지 가시화할 수 있다. 본 실험에서는 임계값으로는 0.7을 두어 그 이상 값에 해당하는 값만 가시화를 하였다. 그림 6은 "바다", "노을", "하늘", "산"에 대한 결과 값을 보여준다. 영상 영역에서는 높은 결과 값에 대해서 하얗게

그림 7. 이미지와 텍스트에 대한 Attention 결과
Fig. 7. Attention results for images and text

표시하고, 텍스트 영역에서는 빨간 글씨로 표시하였다. 우선 “바다” 부분에서 영상 영역에서는 바다 쪽에서 높은 결과 값이 나왔으며 텍스트 영역에서는 “바다”, “제주도”, “물”에서 높은 결과가 나왔다. 다음으로 “노을” 부분에서 영상 영역에서는 붉게 물든 부분이 높은 결과 값이 나왔으며 텍스트 영역에서는 “가을”, “노을”, “석양”, ”노을빛“에 대해서 높은 결과가 나왔다. 그 다음으로 ”하늘“ 부분에서 영상 영역에서는 하얀 구름부분에서 높은 결과 값이 나왔으며 텍스트 영역에서는 ”하늘“, ”구름“, ”에쁘다“에서 높은 결과 값이 나왔다. 마지막으로 ”산“ 부분에서 영상 영역에서는 산 쪽에서 높은 결과 값이 나왔으며 텍스트 영역에서는 ”산“, ”나무“, ”자연“ 부분에서 높은 결과 값이 나왔다. 4가지 부분에서 나온 주의 결과 값을 고려할 때 영상과 텍스트 부분 모두 합리적인 결과가 나타남을 볼 수 있으며, 이는 본 연구에서 수집된 데이터가 MCSM과 같은 대규모 네트워크의 학습에도 과적합 등의 문제를 야기하지 않음을 보여주고 있다.

V. 결론 및 향후과제

본 논문에서는 인스타그램과 같은 SNS를 활용하여 교차양식 기계학습 연구를 위한 대규모 데이터셋을 구축하는 방법을 제안하고, 272,400개의 영상-텍스트 쌍의 데이터셋을 구축하였다. 이러한 교차양식 데이터셋은 영상에 대한 합리적인 품질의 한국어 텍스트로 구성되며, 영상-텍스트 쌍 데이터의 사전 처리를 통해 기계학습에 필요한 특징을 추출하고, 이 특징을 이용해 다양한 딥러닝 모델에서 신속한 학습을 수행할 수 있도록 한다. 이 방법으로 수집한 데이터셋은 인스타그램 이용자의 연령을 고려할 때 가장 현재성이 있기 때문에 교차양식 기계학습에 적합한 한국어 텍스트와 영상의 데이터 쌍이라 말할 수 있다.

또한, 본 논문에서는 구축된 데이터셋의 활용성을 보이기 위해 교차양식 검색 방법을 수행하여 그 결과를 검토하였다. 실험은 전통적인 교차양식 방법, 다양한 DNN 기반의 방법, MCSM 알고리즘 등의 교차양식 검색방법이 수행되었으며, 이들 실험의 결과들로부터 구축된 데이터셋이 비교적 학습 파라메터가 많은 기계학습 모델에서도 과적합 등의 문제를 야기하지 않고 합리적인 교차양식 결과를 제공할 수 있음을 확인하였다.

교차양식 실험결과 MCSM 알고리즘이 가장 높은 성능을 보였으며, 세부 실험을 통해 알고리즘의 성능 및 데이터셋에 대한 검증을 동시에 수행하였다.

인스타그램 데이터셋을 구축할 때, 해당 해시태그와 관련되지 않은 게시물을 자동으로 필터링하였으나 광고 게시물 등의 경우는 한계가 있어 수작업으로 필터링을 해야 하는 어려움이 있었다. 하지만 이러한 게시물들의 특징을 자동으로 찾아낼 수 있다면, 대규모 데이터셋의 구축을 더욱 빠른 속도로 수행 가능할 것이다. 또한, 영상과 텍스트를 한 차원에 임베딩(concatenating)을 할 경우 교차양식 검색뿐만 아니라 교차양식 융합 등 다양한 분야에도 활용할 수 있을 것이며 이러한 후속연구가 진행될 예정이다.

References

[1] J. Deng, W. Dong, R. Socher, L. Jia Li, K. Li, and L. Fei-fei, “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, Jun. 2009.

[2] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ICLR*, 2014.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *NIPS*, pp. 3111-3119, 2013.

[4] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP*, pp. 1746-1751, 2014.

[5] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," ArXiv preprint arXiv:1708.04776, 2017.

[6] L. Fei-fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 28, no. 4, pp. 594-611, Apr. 2006.

[7] G. Griffin, A. Holub, and P. Perona, "*Caltech-256 object category dataset*," Technical Report 7694, California Institute of Technology, 2007.

[8] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object," *ECCV*, pp. 1-15, 2006.

[9] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances NIPS*, pp. 1097-1105, 2012.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ICML*, pp. 448-456, 2015.

[12] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," *ACM-MM*, pp. 251-260, 2010.

[13] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139-147, 2010.

[14] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *IEEE TCSVT*, 2017.

[15] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321-377, 1936.

[16] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE TPAMI*, vol. 36, no. 3, pp. 521-535, 2014.

[17] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *IJCV*, vol. 106, no. 2, pp. 210-233, 2014.

[18] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," *IEEE ICCV*, pp. 4094-4102, 2015.

[19] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," *ACM-MM*, pp. 604-611, 2003.

[20] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semisupervised learning on large graphs," *COLT*, pp. 624-638, 2004.

[21] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," *AAAI*, pp. 1198-1204, 2013.

[22] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semi-supervised regularization," *IEEE TCSVT*, vol. 24, pp. 965-978, 2014.

[23] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised crossmedia feature learning with unified patch graph regularization," *IEEE TCSVT*, vol. 26, no. 3, pp. 583-596, 2016.

[24] Z. Wu, Y. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," *ACM-MM*, pp. 791-800, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770-778, 2016.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep

learning," *ICML*, pp. 689-696, 2011.
[27] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," *ICML*, 2012.
[28] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," *ACMMM*, pp. 7-16, 2014.
[29] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," *ICML*, pp. 1247-1255, 2013.
[30] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," *IJCAI*, pp. 3846-3853, 2016.
[31] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE TCYB*, vol. 47, no. 2, pp. 449-460, 2017.
[32] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE TMM*, vol. 18, no. 7, pp. 1363-1377, 2016.
[33] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *NIPS*, pp. 2204-2212, 2014.
[34] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," *ICML*, pp. 1462-1471, 2015.
[35] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," *CVPR*, pp. 21-29, 2016.
[36] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, and P. Blunsom, "Reasoning about entailment with neural attention," *ICLR*, 2016.
[37] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *NIPS*, pp. 1693-1701, 2015.
[38] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *EMNLP*, pp. 379-389, 2015.
[39] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639-2664, 2004.
[40] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," *ECCV*, pp. 686-701, 2018.
[41] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," *CVPR*, pp. 3128-3137, 2015.
[42] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," in *Proc. 26th Annu. Conf. Human and Cognitive Lang. Technol.*, pp. 133-136, 2014.

김 강 섭 (KangSub Kim)

2017년 2월 : 전북대학교 컴퓨터공학과 공학사
2019년 2월 : 전북대학교 전자공학과 공학석사
<관심분야> 영상처리, 크로스 모달 학습, 인공지능

[ORCID:0000-0003-2403-3811]

이 준 환 (Joonwhoan Lee)

1980년 2월 : 한양대학교 전자공학과 공학사
1982년 2월 : 한국과학기술원 전자공학과 공학석사
1990년 8월 : 미국 미주리대학 전기 및 컴퓨터공학과 공학박사
1990년 10월~현재 : 전북대학교 컴퓨터공학부 교수
<관심분야> 영상처리, 감성 분석, 인공지능

[ORCID:0000-0003-1854-9643]