ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 27, No. 12, pp. 578-583, 2021. 12
https://doi.org/10.5626/KTCP.2021.27.12.578

이미지를 매개로 하는 멀티모달 반지도학습

(Image-Bridged Multimodal Half-Supervised Learning)

김 규 연 † 윤 성 의 ††
(Kyuyeon Kim) (Sung-Eui Yoon)

요 약 멀티모달 데이터를 활용하는 학습 방법은 다양한 형태로 존재하는 데이터를 서로 연관 지어 상호검색을 위한 특징을 추출하거나, 다양한 형태의 데이터를 종합적으로 요구하는 새로운 태스크를 수행하기 위해 사용된다. 현재까지, 이미지와 텍스트 및 이미지와 소리 데이터 간의 멀티모달 학습을 수행하는 연구가 진행되어왔다. 이에 더 나아가, 본 논문에서는 이미지를 중심으로 소리 및 텍스트 데이터를 상호 고려하는 반지도학습 방법을 적용한 모델을 제시한다. 해당 모델은 이미지, 소리, 텍스트를 자유로이 수용하여 각각에 대한 특징을 추출할 수 있다. 덧붙여, 멀티모달 학습에 통상적으로 사용되는 단순 랭킹 손실함수의 한계점을 보완한, 마진값이 이미지 피처 간 유사도에 따라 변하는 가변 마진 랭킹 손실함수를 적용하여 모델을 학습시킨다. 최종적으로, 위 방법을 통해 학습한 모델의 표현력을 평가하기 위해, 제로-샷 텍스트-비디오 검색 성능을 중심으로 이종 데이터 간 상호검색 성능을 정량적으로 분석한다.

키워드: 딥 러닝, 멀티모달 학습, 반지도학습, 가변 마진 랭킹 손실함수, 데이터 상호검색

Abstract Multimodal learning methods have been widely used to extract features from different modalities of data, targeting cross-modal retrieval or accomplishing novel tasks that require a comprehensive understanding of diverse modalities of data. Recent studies have focused on multimodal learning methods by correlating image-text or image-audio. In this paper, we propose a half-supervised learning procedure that alternatively feeds image-audio and image-text pairs. With this training strategy, the model can ingest and extract features from all the given modalities; image, text, and audio, using an image as a bridge. Furthermore, to overcome the limitation of vanilla ranking loss, we propose versatile margin ranking loss that scales the margin considering the similarity between image features. To evaluate the model's representation quality with the proposed strategy, we analyze the quantitative results of cross-modal retrieval between different modalities, primarily focusing on zero-shot text-video retrieval.

Keywords: deep learning, multimodal learning, half-supervised learning, versatile margin ranking loss, cross-modal retrieval

· 본 연구는 삼성전자의 지원(IO201210-07984-01)을 받아 수행된 결과임
· 이 논문은 2020 한국소프트웨어종합학술대회에서 '이미지를 매개로 한 멀티모달 반지도학습 모델'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 한국과학기술원 전산학부 학생
kyuyeonpooh@gmail.com
†† 종신회원 : 한국과학기술원 전산학부 교수(KAIST)
sungeui@kaist.edu
(Corresponding author임)

논문접수 : 2021년 3월 26일
(Received 26 March 2021)
심사완료 : 2021년 10월 13일
(Accepted 13 October 2021)

정보과학회 컴퓨팅의 실제 논문지 제27권 제12호(2021. 12)

1. 서 론

인간은 주어진 정보를 이해하기 위해 다양한 감각 정보를 활용한다. 사람의 말을 이해할 때 말소리와 함께 입 모양에 주목하고, 누군가의 목소리를 들으면 그 사람의 모습을 연상할 수 있는 것처럼, 인간은 한 가지 이상의 감각 정보를 통해 주어진 정보를 성공적으로 해석한다[1].

위와 같은 사실에 근거하여, 멀티모달 학습은 다양한 형태(modality)의 데이터를 함께 학습하는 것이 목표 태스크를 성공적으로 수행하는 데에 도움이 될 수 있다는 직관에 기반한다. 멀티미디어 데이터는 크게 이미지, 소리, 텍스트, 세 가지로 분류된다. 언급된 세 가지 데이터에 대한 멀티모달 학습 방식은 두 가지 독립된 방향으로 발전되어 왔다. 하나는 이미지와 텍스트를 사용하는 멀티모달 학습이다. 이는 이미지와 이에 라벨링 되어 있는 문장 데이터를 활용하는 지도학습(supervised learning) 방식이 널리 채택되고 있으며, 이미지 캡셔닝, 이미지-텍스트 상호검색과 같은 형태로 발전되었다[2]. 다른 하나는 이미지와 소리 간의 멀티모달 학습 방식이다. 연속적 이미지와 소리의 복합체인 비디오 데이터를 활용하는 자기지도학습(self-supervised learning) 방식이 주축을 이루고 있으며, 이미지 내 소리 발생 지점 예측이나 시각 정보를 이용한 소리 발생체 구분 및 소리 분리와 같은 태스크를 목표로 연구가 이루어졌다[3].

본 논문에서는 이미지, 소리, 텍스트, 세 가지 데이터를 자유로이 수용하여, 이들에 대한 저차원적 특징(feature)을 추출할 수 있는 딥 러닝 모델을 제시한다. 그림 1과 같이, 라벨링이 포함된 이미지-텍스트 쌍으로 해당 모델에 속하는 이미지와 텍스트 담당 네트워크를 지도학습시키고, 비디오로부터 이미지와 소리를 추출하여 이미지 및 오디오 네트워크를 자기지도학습 방식으로 훈련시킨다. 이와 같은 과정을 이미지를 매개로 하는 반지도학습(half-supervised learning)이라고 부르기로 한다. 한편, 해당 모델 훈련 시에는 이미지 피처 간의 유사도를 통해 가변적인 마진을 주는 방식으로 랭킹 손실함수(ranking loss)를 변형하여 사용함으로써 기존 랭킹 손실함수의 한계를 완화한다.

본 논문의 이전 연구에서 한 걸음 더 나아가, 텍스트-비디오 검색뿐만 아니라 이미지, 소리, 텍스트 사이의 다양한 경우에 대한 상호검색 성능을 분석한다. 결론적으로, 반지도학습 전략 및 가변 마진 랭킹 손실함수가 전반적인 멀티모달 피처 표현력을 향상하는 것에 기여할 수 있음을 보이고자 한다.

본 논문의 구성은 다음과 같다. 먼저, 2절에서 이미지, 소리, 텍스트를 종합적으로 학습에 활용하는 멀티모달 학습의 관련 연구를 소개한다. 3절에서는 모델 구조, 학습 데이터 구축 방법과 함께, 사용된 손실함수에 대해 설명한다. 4절에서는 학습된 모델로부터 추출된 특징을 활용하여 다양한 이종 데이터 간 상호검색 성능을 정량적으로 분석하고, 5절에서 결론을 서술하며 마친다.

2. 관련 연구

이미지, 오디오, 텍스트를 종합적으로 활용하여 각각에 대한 딥 러닝 피처를 추출하는 연구는 [4]에서 비롯되었다. [4]는 본 논문과 같이 세 가지 형식의 데이터를 저차원의 피처로 표현하는 네트워크를 학습시키고, 추출된 저차원 피처를 통한 이종 데이터 간의 검색 성능을 측정한다. 차이점으로는, 서로 다른 형식에 대한 데이터 피처들이 같은 차원의 공간 내에 서로 알맞게 조정(align)되도록 말단의 fully connected 레이어는 이미지, 오디오, 텍스트 네트워크가 공유하도록 했으며, 모델의 학습을 위해 단순 랭킹 손실함수에 이종 피처 간의 KL 발산 값을 합한 형태의 손실함수를 적용하였다.

[5]에서는 HowTo100M[6]과 AudioSet[7]이라는 대용량 비디오 데이터셋을 활용하여 이미지, 소리, 텍스트 데

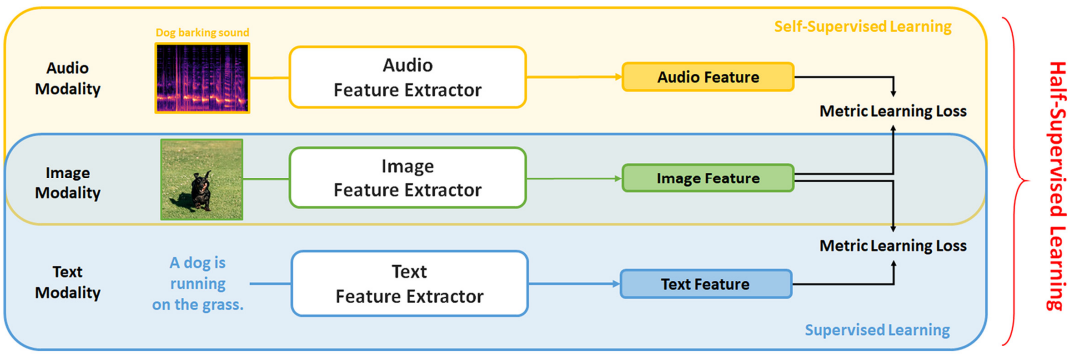

그림 1 이미지를 매개로 하는 반지도학습의 전체적 프로세스
Fig. 1 Overall process of image-bridged half-supervised learning

이터를 동시에 학습하는 모델을 선보였다. 텍스트 데이터를 축약한 임베딩(embedding)은 이미지, 소리 임베딩보다 비교적 낮은 차원 수로 나타낼 수 있다는 가정을 두었으며, 텍스트 데이터를 생성하기 위해 비디오에서 내레이션 음성을 추출한 후, 음성 인식 네트워크를 통해 텍스트 데이터를 얻어내었다. 또한, 더 나은 학습 결과를 위해 이종 데이터 형식 간의 대조 손실함수(contrastive loss)를 새로이 설계하여 학습에 사용하였다.

3. 실험 방법

3.1 데이터 구축

먼저, 이미지와 오디오 네트워크의 학습에 사용될 비디오 데이터를 구축하기 위해 VGGSound[8]를 사용하였다. VGGSound는 약 300종류의 오디오 이벤트가 담긴 19만여 개의 10초 길이 비디오 클립을 제공한다. VGGSound 데이터로부터 비디오 프레임과 소리 데이터를 추출하여, 이로부터 만들어진 이미지-소리 쌍을 모델에 제공함으로써 시각적, 청각적 데이터 간 유기성을 파악할 수 있도록 한다.

비디오 프레임은 1 fps로 추출되어 하나의 비디오 클립당 10개의 이미지를 얻는다. 그림 2의 모델에 속하는 이미지 및 오디오 네트워크 학습에 사용되는 이미지-소리 데이터 쌍은 비디오 클립에서 임의로 고른 하나의 프레임과 해당 프레임을 중간 시점으로 하는 3초 길이의 소리로 구성된다. 소리 데이터는 모노 채널 형태로 추출되어 44,100Hz로 샘플링되었으며, 시간에 따른 소리의 진동수 패턴을 파악하는 방향으로 오디오 네트워크가 학습될 수 있도록 멜-스펙트로그램(mel-spectrogram)으로 전처리하였다. 멜-스펙트로그램 생성 시, 윈도우 크기는 0.1초, hop 크기는 0.05초, mel bin의 수는 128개로 지정하여, 최종적으로 128×301 모양의 소리 데이터를 네트워크에 제공한다. 이미지-소리 데이터 쌍에는 라벨링 과정이 전혀 포함되지 않으며, 오직 비디오 클립 내 동일 구간에 대한 이미지와 소리를 추출하여 쌍을 만드는 방법으로 데이터를 구축하였다.

한편, 이미지-텍스트 데이터셋을 구성하기 위해 COCO의 이미지 캡셔닝 데이터셋과 Flick30k 데이터셋을 합쳤다. 합쳐진 데이터셋에는 약 15만여 개의 이미지-텍스트 쌍이 존재한다. 텍스트 데이터에는 다음과 같은 전처리 과정을 가하였다. 먼저 문장 내 불용어(stopword)를 제거하고, 포함되는 단어의 수가 항상 16개가 되도록 문장의 끝을 잘라내거나 패드를 추가하였다. 그 후, GoogleNews를 통해 사전 학습된 word2vec을 사용하여, 각 단어를 300차원의 벡터로 변환하였다. 최종적으로, 이미지 캡션은 16×300 모양의 배열로 변환되어 텍스트 네트워크에 제공된다.

이미지 데이터는 네트워크에 제공되기 전 일반적인 데이터 확장(augmentation) 기법을 적용하였다. 모든 이미지는 처음에 256×256 크기로 변환되어, 224×224 크기 랜덤 크롭(random crop), 임의 좌우 뒤집기, 그리고 색상 지터(jitter) 과정을 거친다.

이미지-소리, 이미지-텍스트 데이터셋에 대해 90%에 해당하는 데이터를 훈련 데이터로 사용하였고, 나머지 10%의 데이터는 검증(validation) 데이터로 사용하였다. 테스트 데이터는 MSR-VTT를 사용하므로 별도로 만들지 않았다. 테스트에 사용되는 모델은 검증 데이터에서 이미지-소리, 이미지-텍스트 상호검색 성능이 평균적으로 가장 좋았던 시점의 모델을 선택하였다.

3.2 모델 구조 및 손실 함수

그림 2와 같이, 실험에 사용된 모델은 이미지, 오디오, 텍스트, 세 가지 부차적인 네트워크로 구성된다. 이미지 네트워크는 ResNet-18를 사용하였으며, ImageNet에 사전 학습된 파라미터로 초기화하였다. 오디오 네트워크 또한 [9]의 결과에 근거하여, 마찬가지로 ResNet-18를 특징 추출 네트워크로 사용하였다. 한편, word2vec으로 전처리된 텍스트는 여러 개의 1D 컨볼루션 레이어를 거

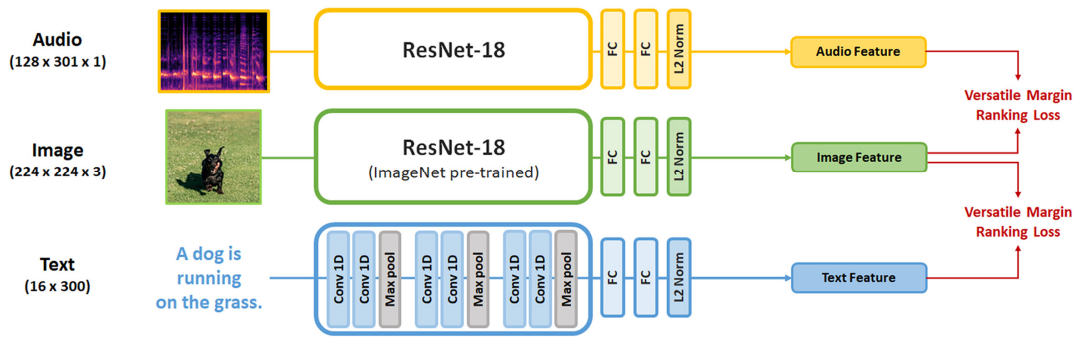

그림 2 모델 구조 및 학습 과정

Fig. 2 Model architecture and its training process

쳐 저차원 피처로 압축된다. 모든 서브 네트워크 내 컨볼루션 레이어에는 배치 정규화(batch normalization) 및 ReLU 레이어가 뒤따른다. 각 서브 네트워크를 통과한 피처는 독립된 두 개의 fully connected 레이어를 통과하여, 마지막으로 L2 정규화(normalize)가 가해진다. 이렇게 얻어낸 최종적인 피처를 상호검색 성능 측정 및 손실(loss) 계산에 이용한다.

멀티모달 학습에 통상적으로 사용하는 단순 랭킹 손실함수의 정의는 수식 (1)과 같다.

$$L_{xy} = \sum_i \sum_{j \neq i} \max(0, x_i^\top y_j - x_i^\top y_i + \alpha) \quad (1)$$

여기서, x는 앵커(anchor)로 쓰이는 한 형식의 데이터를, y는 다른 형식의 데이터를 의미하며, α는 마진을 표기한다. i와 j는 데이터가 배치 내에서 몇 번째에 해당하는지를 나타내므로, $x_i^\top y_j$는 다른 쌍에서 온 피처 간의 유사도를, $x_i^\top y_i$는 쌍을 이루는 데이터에 대한 피처 간 유사도를 의미한다. 따라서, 단순 랭킹 손실함수는 같은 쌍에서 온 피처 간 유사도가 다른 쌍에서 얻은 피처 간 유사도보다 지정된 마진만큼 상대적으로 크게 하는 효과를 준다.

그러나, 단순 랭킹 손실함수는 다른 쌍에서 온 피처 간 유사도가 동일 쌍에서 얻은 피처 간 유사도와 무조건 α만큼 차이가 나도록 한다. 따라서, 다른 쌍에서 온 데이터가 유사성을 띠어도, 유사성을 전혀 띠지 않는 것과 동일한 만큼의 마진이 지정된다. 이러한 점을 보완하기 위해, 본 연구에서는 가변 마진 랭킹 손실함수(versatile margin ranking loss)를 사용했으며, 수식 (2)와 같이 정의한다.

$$L_{xy}^{VM} = \sum_i \sum_{j \neq i} \max(0, x_i^\top y_j - x_i^\top y_i + \sigma(-x_i^\top x_j) \cdot \alpha) \quad (2)$$

σ는 sigmoid 함수를 의미하며, 수식의 $-x_i^\top x_j$은 단순히 마진의 크기를 조절하는 역할을 맡고 있으므로, 이 항에 대한 그래디언트가 발생하지 않도록 하였다 .가변 마진 랭킹 손실함수는 다른 쌍에서 온 앵커 피처 간의 유사도를 고려하여 가변적인 마진 값을 부여한다. x를 이미지, y를 오디오 데이터 형식인 경우를 가정하자. 서로 다른 비디오에서 추출된 두 이미지가 있을 때, 두 이미지의 내용이 서로 유사하다면, $\sigma(-x_i^\top x_j)$ 값은 상대적으로 작으므로, 더 적은 마진이 부여된다. 따라서 각 이미지에 대응되는 오디오가 서로 약하게 밀어내는 효과를 준다. 반면, 두 이미지의 내용이 무관하다면 $\sigma(-x_i^\top x_j)$ 값은 커지고, 상대적으로 큰 마진 값이 지정된다.

최종 손실함수는 단순 랭킹 손실함수와 가변 마진 랭킹 손실함수를 혼합한 형태이다. 안정적인 학습을 위해 가변 마진을 부여하는 방식은 이미지 데이터가 앵커일 경우에만 적용하였으며, 그 외의 경우에는 단순 랭킹 손실함수를 적용하였다. 이는 수식 (3)과 같다.

$$L = L_{va}^{VM} + L_{av} + L_{vt}^{VM} + L_{tv} \quad (3)$$

수식 (3)에서 v는 시각적(visual) 데이터 형식, 즉 이미지를, a는 오디오 데이터를, t는 텍스트 데이터 형식을 의미한다. 즉, 이미지 데이터가 앵커일 경우, 이미지 데이터 간 유사도를 고려해 가변적인 마진을 부여하게 된다.

하이퍼파라미터(hyper-parameter) 설정은 다음과 같다. 모델 훈련 시, 배치 사이즈(batch size)는 32로 설정하였으며, Adam optimizer를 사용하였다. Learning rate는 이미지 네트워크에 대해 5e-5로, 오디오 및 텍스트 네트워크에 대해 1e-4로 설정하였다. 30 epoch 후 learning rate의 크기를 1/10 수준으로 낮추었으며, 총 40 epoch 정도 학습시켜 검증 데이터에 가장 좋은 성능을 보이는 모델을 채택하여 테스트에 사용하였다. 위 학습 과정은 2개의 GTX 1080 Ti GPU를 사용하여 약 하루 정도 소요된다.

4. 결 과

4.1 성능 지표 및 평가 방법

MSR-VTT는 비디오 캡셔닝 데이터셋으로, 비디오 클립과 그에 대응되는 문장 형식의 캡션으로 이루어져 있다. 다양한 데이터 형식에 대한 피처의 표현력을 평가하기 위해 MSR-VTT[10] 데이터셋 중 테스트 스플릿(1,000개 비디오-텍스트 쌍)을 사용하여, 제로-샷 텍스트-비디오 검색 성능을 포함한 타 이종 데이터 간 제로-샷 상호검색 성능까지 폭넓게 분석해보았다.

성능 평가 지표는 R@K (Recall at K)와 MedR을 사용한다. R@K는 쿼리에 대한 K개 추출된 검색 항목 중에서 쿼리와 쌍을 이루는 항목이 존재하는지 판단하는 지표이다. 즉, R@K의 값이 클수록 모든 쿼리에 대해 원하는 항목을 성공적으로 검색한 비율이 높음을 의미한다. 한편, 쿼리에 대한 검색 항목들을 유사도가 큰 순으로 나열했을 때, MedR은 나열된 항목 중 정답 항목에 해당하는 순위(rank)들을 구하고, 모든 쿼리에 대한 그 값의 중간값을 의미한다. 쿼리와 쌍을 이루는 정답 항목의 순위가 높을수록 좋으므로, MedR은 낮을수록 바람직하다.

4.2 결과 분석

표 1은 MSR-VTT의 테스트 스플릿에서 측정한 관련 연구 및 제시한 모델의 텍스트-비디오 검색 성능을 정리한 표이다. 표 1을 기준으로 랜덤 방법을 제외한 위

3개 방법[11-13]은 MSR-VTT 데이터셋에서 학습된 모델들이다. 해당 모델들의 학습 데이터와 테스트 데이터 도메인이 MSR-VTT로 동일하다. 즉, MSR-VTT에 대해 완전한 지도학습 방식으로 훈련되었다. 이들 중 JSFusion[13]이 가장 우수한 성능을 보인다. 본 연구에서 제시한 반지도학습 모델은 훈련 데이터의 도메인이 다름에도 불구하고, 같은 데이터 도메인에 대해 지도학습을 거친 세 모델 중 일부보다 우수한 성능을 나타냄을 확인할 수 있다.

다음 세 가지 모델은[5,6,14] HowTo100M 데이터셋에서 훈련되었다. 표 1에서 알 수 있듯이, HowTo100M은 훈련 데이터에 해당하는 비디오들의 총 시간이 15년을 상회하는 방대한 데이터셋이다. 해당 모델들은 본 연구의 모델보다 좋은 성능을 나타내고 있지만, 훈련에 사용된 학습 데이터의 규모와 이에 요구되는 컴퓨팅 파워를 고려했을 때 제시된 반지도학습 모델도 준수한 검색 성능을 나타내고 있음을 주장할 수 있다.

표 1의 Ours는 단순 랭킹 손실 함수만으로 학습한 모델을 통해 추출한 피처에 대한 검색 성능을 나타낸다. VM-Ranking Loss는 가변 마진 랭킹 손실함수를 지칭하며, 가변 마진을 주는 손실함수를 적용했을 때 전체적인 성능이 개선된 것을 볼 수 있다. 이에 더해, Feature Mix는 검색에 사용될 피처에 다른 데이터 형식의 피처를 일정 비율 혼합시킨 것을 검색에 사용하였음을 의미한다. 표 1의 경우 검색될 비디오에 오디오 피처를 일부 혼합하여 검색에 활용하였다. 이 결과, 복합 피처를 검색 수단으로 활용한 결과, 성능이 더 개선된 것을 알 수 있다.

표 2에서는 텍스트-비디오 검색 외 다양한 데이터 형식 간 상호검색 성능을 나열하고 있다. Query, Retrieve는 각각 쿼리 및 검색하고자 하는 데이터 형식을 의미한다. 비교를 위해 [4]의 모델을 가져와 본 연구에서 구축한 데이터셋을 학습시켜 성능을 측정하였다. 본 연구

표 1 MSR-VTT test-split에서의 텍스트-비디오 검색 성능
Table 1 Performance of Text-to-Video retrieval in MSR-VTT test-split

Method	Train Data	Video Amount	Learning	R@1	R@5	R@10	MedR
Random	-	-	-	0.1	0.5	1.0	500
Torabi et al. [11]	MSR-VTT	41.2 hours	Supervised	4.2	12.9	19.9	55
VSE-LSTM [12]				3.8	12.7	17.1	66
JSFusion [13]				10.2	31.2	43.2	13
Miech et al. [6]	HowTo100M (+AudioSet)	15 years	Self-Supervised	7.5	21.2	29.6	38
MILNCE [14]				9.9	24.0	32.4	30
MMV FAC [5]		15(+1) years		9.3	23.0	31.1	38
Ours	VGGSound +COCO +Flickr30k	550+ hours	Half-Supervised	5.7	16.7	24.3	49
w/ VM Ranking Loss				6.3	**17.1**	25.1	48
+ w/ Feature Mix (Audio)				**6.9**	17.0	**25.4**	**44**

표 2 MSR-VTT test-split에서의 다양한 이종 데이터간 상호검색 성능
Table 2 Performance of diverse cross-modal retrieval in MSR-VTT test-split

Query	Retrieve	Method	R@1	R@5	R@10	MedR
-	-	Random	0.1	0.5	1.0	500
Image (Video)	Audio	See, Hear, and Read [4]	2.1	8.9	13.8	127
		Ours (w/ Vanilla Ranking Loss)	2.7	9.6	15.0	112
		Ours (w/ VM Ranking Loss + Feat. Mix)	**2.8**	**10.0**	**16.0**	**101**
Image (Video)	Text	See, Hear, and Read [4]	6.1	15.3	21.6	58
		Ours (w/ Vanilla Ranking Loss)	**6.7**	**16.6**	24.7	**48**
		Ours (w/ VM Ranking Loss + Feat. Mix)	6.5	**16.6**	**25.9**	**48**
Audio	Image	See, Hear, and Read [4]	2.3	8.1	12.9	108
		Ours (w/ Vanilla Ranking Loss)	1.9	8.3	13.5	109
		Ours (w/ VM Ranking Loss + Feat. Mix)	**2.5**	**9.1**	**13.7**	**102**
Audio	Text	See, Hear, and Read [4]	1.0	3.3	6.9	189
		Ours (w/ Vanilla Ranking Loss)	0.9	4.1	7.4	171
		Ours (w/ VM Ranking Loss + Feat. Mix)	**1.2**	**4.8**	**7.9**	**163**
Text	Audio	See, Hear, and Read [4]	1.3	5.1	8.9	167
		Ours (w/ Vanilla Ranking Loss)	1.3	5.0	9.1	162
		Ours (w/ VM Ranking Loss + Feat. Mix)	**1.4**	**5.9**	**9.8**	**151**

에서 제시한 반지도학습 기반 모델이 쿼리, 검색 데이터 형식과 무관하게 [4]의 모델보다 두루 좋은 성능을 나타내고 있다. 또한, 단순 랭킹 손실함수를 통해 학습한 모델보다 가변 마진 랭킹 손실함수와 피처 혼합 전략을 적용한 모델이 성능 개선에 기여하고 있음을 알 수 있다.

5. 결 론

본 논문은 이미지-소리 및 이미지-텍스트 데이터셋을 이용해, 이미지를 매개로 하여 서로 다른 3가지 형식의 데이터를 학습하는 멀티모달 반지도학습과 이를 적용한 모델을 제시한다. 제시한 모델은 이미지, 소리, 텍스트 데이터를 모두 수용할 수 있고, 각각을 저차원의 피처로 표현하는 방법을 학습한다. 학습 시 가변적인 마진을 주는 랭킹 손실함수를 사용하여 기존 랭킹 손실함수의 한계점을 보완하였다. 결과적으로, 이종 피처 간 상호검색을 수행했을 때, 가변 마진 손실함수 적용한 경우 또는 이종 데이터 피처를 혼합한 경우 성능이 개선되는 것을 확인할 수 있었다.

여러 형식의 데이터에 대한 멀티모달 학습을 수행할 때, 모든 형식에 대해 싱크를 이루는 데이터셋을 구하는 것은 어렵다. 본 논문은, 학습시키고자 할 다양한 형식(이미지, 소리, 텍스트)의 데이터가 모두 상응하지 않더라도, 특정 형식(이미지)을 기준점으로 쌍을 이루는 데이터(이미지-소리, 이미지-텍스트)를 구할 수 있다면, 주어진 모든 데이터 형식들에 대한 종합적 멀티모달 학습이 가능할 수 있음을 시사하고 있다.

References

[1] R. Arandjelovic and A. Zisserman, "Objects that Sound," *ECCV*, 2018.

[2] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked Cross Attention for Image-Text Matching," *ECCV*, 2018.

[3] A. Owens and A. A. Efros, "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features," *ECCV*, 2018.

[4] Y. Aytar, C. Vondrick, and A. Torralba, "See, Hear, and Read: Deep Aligned Representations," *arXiv*, 2017.

[5] J. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. D. Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-Supervised MultiModal Versatile Networks," *NeurIPS*, 2020.

[6] A. Miech, D. Zhukov, J. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," *ICCV*, 2019.

[7] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," *ICASSP*, 2017.

[8] H. Chen, W. Xie, A. Vedaldi, A. Zisserman, "VGG-Sound: A Large-Scale Audio-Visual Dataset," *ICASSP*, 2020.

[9] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking CNN Models for Audio Classification," *arXiv*, 2020.

[10] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," *CVPR*, 2016.

[11] A. Torabi, N. Tandon, and L. Sigal, "Learning Language-Visual Embedding for Movie Understanding with Natural-Language," *arXiv*, 2016.

[12] Y. Yu, H. Ko, J. Choi, and G. Kim, "Video Captioning and Retrieval Models with Semantic Attention," *arXiv*, 2016.

[13] Y. Yu, J. Kim, and G. Kim, "A Joint Sequence Fusion Model for Video Question Answering and Retrieval," *ECCV*, 2018.

[14] A. Miech, J. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-End Learning of Visual Representations from Uncurated Instructional Videos," *CVPR*, 2020.

김 규 연

2020년 성균관대학교 소프트웨어학과(학사). 2020년~현재 한국과학기술원 전산학부 석사과정 재학중. 관심분야는 멀티모달 학습, 시청각 학습, 비디오 인페인팅

윤 성 의

1999년 서울대학교 Computer Science(학사). 2001년 서울대학교 Computer Science(석사). 2005년 University of North Carolina at Chapel Hill(박사). 2005년~2007년 Lawrence Livermore National Laboratory, USA. 2007년~현재 한국과학기술원 전산학부 교수. 관심분야는 렌더링, 이미지 검색, 컴퓨터 비전, 로보틱스