이미지를 매개로 한 멀티모달 반지도학습 모델

김규연○, 윤성의
한국과학기술원 전산학부
kyuyeonpooh@kaist.ac.kr, sungeui@kaist.edu

Image-Bridged Multimodal Half-Supervised Learning Model

Kyuyeon Kim○, Sung-Eui Yoon
School of Computing, Korea Advanced Institute of Science and Technology

요 약

멀티모달 데이터를 사용하는 학습 방법은 다양한 형태로 존재하는 데이터를 서로 연관지어, 상호 검색을 위한 특징을 추출하거나, 종합적 형태의 데이터를 요구하는 새로운 태스크를 수행하기 위해 사용된다. 현재까지 이미지와 텍스트 및 이미지와 소리 데이터 간의 멀티모달 학습을 수행하는 연구가 진행되어왔다. 이에 더 나아가, 본 논문에서는 이미지, 소리, 텍스트 데이터를 모두 수용할 수 있고, 이를 복합적으로 고려하여 특징을 추출하는 모델을 제시한다. 해당 모델은 이미지-소리 데이터셋과 이미지-텍스트 데이터셋을 학습에 사용하여, 이미지, 소리 및 텍스트 데이터 모두에 대해 총체적인 유기성을 파악하는 것을 목표로 하는 이미지 매개 반지도학습 모델이다. 덧붙여, 해당 모델은 멀티모달 학습에 통상적으로 적용되는 단순 랭킹 손실함수의 한계점을 보완하여, 마진 값이 앵커 피쳐 간 유사도에 따라 달라지는 변형된 형태의 랭킹 손실함수를 통해 학습한다. 최종적으로, 다양한 형태의 데이터에 대한 모델의 표현력을 평가하기 위해 텍스트-비디오 검색 성능 결과를 제시하여, 위 방법으로 학습한 모델의 성능을 확인한다.

1. 서 론

인간은 주어진 정보를 이해하기 위해 한 가지 감각만을 이용하지 않는다. 사람의 말을 이해할 때 소리와 함께 입 모양을 참고하고, 누군가의 목소리를 들을 때 그 사람의 얼굴을 연상하는 것처럼, 인간은 한 가지 이상의 감각 정보를 통해 주어진 정보를 성공적으로 해석한다[1].

위와 같은 사실에 입각하여, 멀티모달 학습은 다양한 형태의 데이터를 함께 학습하는 것이 목표 태스크를 성공적으로 수행하는 데에 도움이 될 것이라는 직관에 기반한다. 멀티미디어 데이터는 크게 이미지, 소리, 텍스트, 세 가지로 분류할 수 있다. 언급된 세 가지 데이터에 대한 멀티모달 학습 방식은 두 가지 방향으로 발전되어 왔다. 하나는 이미지와 텍스트를 사용하는 멀티모달 학습이며, 이는 이미지와 이에 라벨링 되어있는 문장 데이터를 활용하는 지도학습(supervised learning) 방식을 띠고 있으며, 이미지 캡셔닝, 이미지-텍스트 상호 검색과 같은 형태로 발전하였다[2]. 다른 하나는 이미지와 소리 간의 멀티모달 학습으로, 연속적 이미지와 소리의 복합체인 비디오 데이터를 활용하는 자기지도학습(self-supervised learning) 방식이 주축을 이뤘고, 이미지 내 소리 발생 지점 예측이나 시각 정보를 이용한 소리 분리와 같은 태스크를 목표로 연구가 되어왔다[3].

본 논문에서는 이미지, 소리, 텍스트, 세 가지 데이터를 모두 수용하여, 이들에 대한 저차원적 특징을 추출할 수 있는 딥 러닝 모델을 제시한다. 라벨링이 포함된 이미지-텍스트 쌍으로 이미지와 텍스트 담당 네트워크를 지도학습시키고, 비디오로부터 이미지와 소리를 추출하여 이미지 및 오디오 네트워크를 자기지도학습 방식으로 훈련시키기 때문에, 이를 이미지를 매개로 하는 반지도학습(half-supervised learning)이라고 부르기로 한다. 한편, 해당 모델 훈련시에는 앵커(anchor) 피쳐 간의 유사도가 고려된 가변 마진을 주는 방식으로 랭킹 손실함수를 변형하여 사용함으로써 기존 랭킹 손실함수의 한계를 완화한다.

본 논문의 구성은 다음과 같다. 먼저, 2절에서 이미지, 소리, 텍스트를 종합적으로 이용하는 멀티모달 학습에 대한 관련 연구를 소개한다. 3절에서는 모델 구조, 학습 데이터 구축 방법과 함께, 사용된 손실함수에 대해 설명한다. 4절에서는 학습된 모델로부터 추출된 특징으로 비디오-텍스트 상호 검색을 통해 성능을 제시하고, 5절에서 결론을 서술한다.

2. 관련 연구

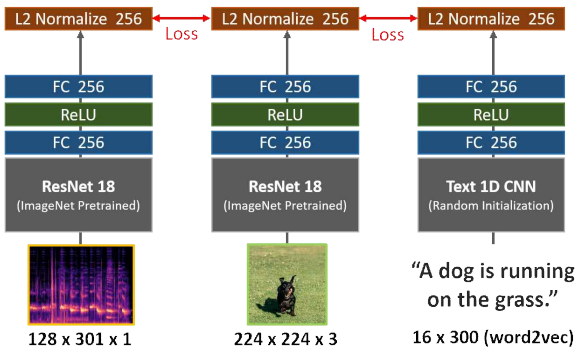
이미지, 오디오, 텍스트를 종합적으로 활용하고 이들의 딥 러닝 피쳐를 추출하는 연구는 [4]에서 비롯되었다. [4]는 본 논문과 같이 세 가지 형식의 데이터를 저차원의 피쳐로 표현하는 네트워크를 학습시키고, 추출된 저차원 피쳐를 통한 이종 데이터 간의 검색 성능을 측정한다. 차이점으로는, 서로 다른 형식의 데이터 피쳐가 서로 잘 조정(align)되도록 말단의 fully conncted 레이어는 이미지, 오디오, 텍스트 네트워크가 공유하도록 했으며, 단순 랭킹 손실함수에 이종 피쳐 간의 KL 발산을 합한 형태의 손실함수를 사용하였다. 해당 연구는 모델 성능 평가 시 공개 데이터를 사용하지 않아 성능 비교가 어렵다.

[5]에서는 HowTo100M[6]과 AudioSet[7]이라는 대용량 비디오 데이터셋을 활용하여 이미지, 소리, 텍스트 데이터를 동시에 학습하는 모델을 선보였다. 텍스트 데이터를 나타내는 임베딩은 이미지, 소리 임베딩에 비해 낮은 차원을 가지는 것이 더 바람직할 것이라는 가정을 둔 것과, 텍스트 데이터를 만들기

* 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구입니다. (No. NRF-2017M3C4A7066317).

위해 영상 내 내레이션을 음성인식 네트워크를 통해 텍스트로 변환했다는 점이 특징이다. 서로 다른 데이터 형식 간에 대조(contrastive) 손실함수를 정의하여 학습에 사용하였다.

3. 실험 방법



[그림 1] 모델 구조도.

3.1. 데이터 구축

먼저, 이미지와 오디오 네트워크의 학습을 위한 비디오 데이터 구축하기 위해 VGGSound[8]를 사용하였다. VGGSound는 약 300종류의 오디오 이벤트가 담긴 19만여개의 10초 길이 비디오 클립을 제공한다. 제시 모델은 VGGSound 데이터로부터 추출한 비디오 프레임과 소리 데이터를 연관지어 학습함으로써 시각적, 청각적 데이터 간의 유기성을 파악할 수 있다. 비디오 프레임은 1 fps로 추출하여 한 비디오 클립 당 10개의 이미지를 얻었다. 그림 1의 이미지, 오디오 네트워크 학습에 사용되는 이미지-소리 데이터 쌍은 비디오 클립에서 랜덤하게 고른 하나의 프레임과, 해당 프레임을 중간 시점으로 하는 3초 길이의 오디오로 구성하였다. 오디오 데이터는, 시간에 따른 소리의 진동수를 네트워크가 잘 파악할 수 있도록 멜-스펙트로그램으로 변환하여 학습에 사용하였다. 이미지-오디오 데이터 쌍에는 라벨링 과정이 포함되지 않으며, 비디오 내 같은 구간에 대한 프레임과 소리를 추출하는 방법으로 데이터를 구축하였다.

한편, 이미지-텍스트 데이터셋 마련을 위해 COCO의 이미지 캡셔닝 데이터셋과 Flick30k 데이터셋을 합쳤다. 합쳐진 데이터셋은 약 15만여개의 이미지-텍스트 쌍으로 이루어져 있다. 학습에 사용하기 위해 몇 가지 전처리 과정을 거쳤다. 먼저 문장 내 불용어(stopword)를 제거하고, 항상 16개 단어가 포함되도록 문장의 끝을 잘라내거나 패딩하였다. 그 후, GoogleNews의 word2vec을 사용하여 각 단어를 300차원의 벡터로 변환하였다.

3.2. 모델 구조 및 손실 함수

그림 1에서 볼 수 있듯이, 실험에 사용된 모델은 이미지, 오디오, 텍스트의 세 가지 부차적인 네트워크로 구성된다. 이미지 네트워크는 ResNet-18를 사용하였으며, ImageNet에 사전 학습된 웨이트(weight)로 초기화하였다. 오디오 네트워크의 경우도 ImageNet 사전 학습된 네트워크가 소리 특징 추출에 도움이 된다는 [9]의 결과에 입각하여, 마찬가지의 ResNet-18를 사용하였다. 한편, word2vec으로 전처리된 텍스트는 1D CNN을 거치게 하여 피쳐를 얻는다. 각 네트워크의 말단에는 두 개의 fully connected 레이어가 있으며, 최종적으로 L2 정규화를 가하여 각 데이터에 대한 피쳐를 구한다.

멀티모달 학습에 통상적으로 사용하는 단순 랭킹 손실함수는 다음과 같다.

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

여기서, x는 앵커(anchor)로 쓰이는 한 형식(modality)의 데이터를, y는 다른 형식의 데이터를 의미하며, α는 마진을 표기한다. i와 j는 데이터가 몇 번째인지를 의미하므로, $x_i^\top y_j$는 다른 쌍에서 온 피쳐 간의 유사도를, $x_i^\top y_i$는 쌍을 이루는 데이터에 대한 피쳐 간 유사도를 지칭한다. 따라서, 단순 랭킹 손실함수는 같은 쌍에서 온 피쳐 간 유사도가 다른 쌍에서 얻은 피쳐 간 유사도보다 상대적으로 높도록 만드는 효과를 준다.

그러나, 단순 랭킹 손실함수는 다른 쌍에서 온 피쳐 간 유사도가 동일 쌍에서 얻은 피쳐 간 유사도와 무조건적으로 α만큼 차이가 나도록 한다. 따라서, 다른 쌍에서 온 데이터가 상당히 유사할 수 있음에도 항상 동일한 마진이 설정된다. 이러한 점을 보완하기 위해, 본 연구에서는 가변 마진 랭킹 손실함수를 사용했으며, 다음과 같이 정의한다.

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

σ는 sigmoid 함수를 의미하며, 이 손실함수는 다른 쌍에서 온 앵커 피쳐 간의 유사도를 고려하여 가변적인 마진 값을 부여한다. x를 이미지, y를 오디오라고 가정하면, 이미지와 오디오가 서로 다른 비디오에서 왔어도, 앵커인 이미지 피쳐 간의 유사도가 높으면 상대적으로 낮은 마진이, 반대의 경우에는 비교적 높은 마진이 적용된다. 덧붙여, 수식의 $-x_i^\top x_j$은 단순히 마진의 크기를 조절하는 역할을 맡고있기 때문에, 이 항에 대한 그래디언트가 발생하지 않도록 하였다.

최종 손실함수는 단순 랭킹 손실함수와 가변 마진 랭킹 손실함수를 혼합한 형태이다. 가변 마진 랭킹 손실 함수는 이미지 데이터가 앵커일 경우에만 사용하였으며, 그 외의 경우에는 단순 랭킹 손실함수를 적용하였다. 이는 다음과 같다.

$$L = L_{va}^V + L_{av} + L_{vt}^V + L_{tv}$$

위 식에서 v는 시각적(visual) 데이터 형식, 즉 이미지를, a는 오디오 데이터를, t는 텍스트 데이터 형식을 의미한다.

모델 훈련시 batch 크기는 512, 옵티마이저는 Adam을 사용하였다. 학습률(learning rate)은 초기 1e-4로 설정하였고, 30 epoch 뒤에는 1e-5로 낮추었으며, 약 40 epoch 동안 학습시켜 검증 데이터셋에 가장 좋은 성능을 보이는 모델을 채택했다. 위 과정은 2개의 GTX 1080 Ti GPU를 사용하여 약 하루 정도가 소요되었다.

4. 결과

제시한 모델의 다양한 데이터 형식에 대한 피쳐의 표현력을 평가하기 위해 MSR-VTT[10] 데이터셋 중 [5, 6, 14]에 쓰인 테스트 스플릿을 사용하여, 제로-샷(Zero-shot) 텍스트-비디오 검색 성능을 측정하였다. 성능 평가 지표는 R@K(Recall at K)와 MedR을 사용한다. R@K는 K개 추출된 항목 중 쿼리와 완

[표 1] MSR-VTT 데이터셋 (테스트 스플릿) 에서의 Text-to-Video Retrieval 성능.

Method	Train Data	Video Amount	Learning	R@1	R@5	R@10	MedR
Random	-	-	-	0.1	0.5	1.0	500
C+LSTM+SA+FC7[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]		15 years		**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	5.3	16.3	23.6	52
+Audio Feature				6.0	**16.6**	**25.0**	49
+VM-Ranking Loss				**6.1**	15.8	24.2	**48**

벽히 대응되는 아이템의 비율을, MedR은 추출 항목 중 쿼리와 대응되는 항목이 위치한 순위들의 중간값을 의미한다.

표 1은 MSR-VTT 데이터셋에서 텍스트를 통한 비디오 검색 성능을 정리한 표이다. 랜덤 방법을 제외한 위 3개 방법[11, 12, 13]은 MSR-VTT 데이터셋으로 학습된 모델들의 결과이다. MSR-VTT는 라벨링이 포함되었기 때문에, 해당 모델들은 지도학습으로 훈련되었다. JSFusion[13]이 가장 우수한 성능을 보였지만, 본 연구에서 제시한 반지도학습 모델이 지도학습을 거친 일부 모델보다 우수한 성능을 나타냄을 확인할 수 있다. 다음 3개[5, 6, 14]는 HowTo100M 데이터셋을 학습한 모델들이다. 표 1에서 알 수 있듯이, HowTo100M은 비디오들의 총 시간이 15년을 상회하는 방대한 데이터셋이다. 해당 모델들은 본 연구의 모델보다 높은 성능을 나타내지만, 학습 데이터의 규모를 고려하면 제시 모델도 준수한 검색을 나타냄을 주장할 수 있다.

표 1의 Ours는 단순 랭킹손실 함수로 학습한 모델로 추출한 텍스트와 이미지 피쳐를 통해 얻어낸 검색 성능을 의미한다. Ours 하단의 항목은 비디오 검색시 이미지 피쳐에 소리 피쳐를 일정 비율 섞은 피쳐를 통해 수행된 비디오 검색 성능을 나타낸다. 복합 피쳐를 검색 수단으로 활용한 결과, 성능이 상당히 오른 것을 확인할 수 있다. VM-Ranking Loss는 가변 마진 랭킹 손실함수를 의미하며, 이전 항목보다도 R@1 및 MedR 측면에서 성능이 오른 것을 알 수 있다. 본 논문에서 직접 제시되지는 않았으나, 다른 데이터 형식에 대한 상호 검색 성능을 분석했을 때에도 타종 피쳐를 혼합하여 검색에 사용하는 방법과 가변 마진 손실함수를 사용한 제시 모델이 더 우수한 검색 성능을 보이는 경향성을 확인할 수 있었다.

5. 결론

본 논문은 이미지-소리 및 이미지-텍스트 데이터셋을 이용해, 이미지를 매개로 하여 서로 다른 3가지 형식의 데이터를 학습하는 멀티모달 반지도학습 모델을 제시한다. 제시한 모델은 이미지, 소리, 텍스트 데이터를 받아 저차원의 피쳐로 표현하는 방법을 학습한다. 또한, 기존 랭킹 손실함수의 한계점을 보완한 가변 마진 손실함수를 사용한다. 결과적으로, 텍스트를 통한 비디오 검색 성능을 분석했을 때 준수한 성능을 나타냈으며, 서로 다른 종류의 피쳐를 혼합했을 때와 가변 마진 손실함수를 사용했을 때 성능이 개선되는 것을 확인할 수 있었다.

여러 형식의 데이터에 대한 멀티모달 학습을 수행할 때, 모든 형식에 대해 싱크를 이루는 데이터셋을 구하는 것을 어려운 일이다. 본 논문은, 하나의 데이터 형식(본 논문에서는 이미지)을 기준으로 짝을 이루는 다른 형식의 데이터들이 존재할 때, 기준 형식을 매개로 서로 다른 데이터 형식에 대한 종합적 멀티모달 학습이 가능할 수 있음을 시사한다.

6. 참고 문헌

[1] R. Arandjelovic, 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, 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] HowTo100M, https://www.di.ens.fr/willow/research/howto100m/.

[7] AudioSet, https://research.google.com/audioset/.

[8] VGGSound, http://www.robots.ox.ac.uk/~vgg/data/vggsound/.

[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.