비디오 탐색을 위한 이미지/오디오 기반 유사도 선택 네트워크

윤종수 [1], 오유정 [2], 차수연 [1], 최종원 [1)2)]*
[1)]중앙대학교 AI 학과, [2)]중앙대학교 첨단영상대학원 영상학과
{whyjs, maryoh, suyeon.cha}@vilab.cau.ac.kr, choijw@cau.ac.kr*

Image/Audio based similarity selecting network for video retrieval

Jongsu Youn[1], Yujeong Oh[2], Suyeon Cha[1], Jongwon Choi[1)2)]*
[1)]Department of Artificial Intelligence, Chung-Ang University
[2)]Department of Advanced Imaging, GSAIM, Chung-Ang University

요 약

비디오 탐색은 비디오 검색, 행동 인지, 비디오 추천 등과 같은 컴퓨터 비전 연구에서 가장 중요한 분야이다. 비디오 데이터는 여러 modality 를 포함하고 있다. 텍스트를 활용한 비디오 관련 연구가 많지만, 텍스트와 같은 메타데이터는 별개의 annotation 작업이 요구된다. 본 논문에서는 별도의 annotation 없이 비디오 데이터 자체에 포함된 이미지와 오디오 데이터를 탐색에 활용한다. 이를 위해 이미지/오디오 기반 유사도를 선택적으로 활용하는 multi-modal 비디오 탐색을 위한 네트워크를 제안한다.

Ⅰ. 서론

비디오는 이미지, 오디오, 텍스트 등 다양한 modality 를 포함한 데이터이다. 하지만, 대부분의 비디오 탐색 연구에서는 이미지 데이터를 중점적으로 다루고 있다. 본 논문에서는 여러 modality 정보를 활용하여, 비디오에 대한 더 풍부한 정보를 학습하고, 특정 modality 에 대한 의존성을 개선하고자 한다.

비디오의 여러 modality 중 텍스트 정보는 영상 촬영 이외의 annotation 작업이 추가로 필요하며, 작업 과정에서 생기는 오류들은 딥러닝 기반 task 에서 문제를 일으킬 수 있다. 이에 따라 이미지와 오디오와 같이 비디오 데이터에서 자체적으로 추출할 수 있는 정보들만을 입력으로 받는 비디오 탐색 모델에 관한 연구의 중요도가 높아지고 있다.

본 논문은 메타 데이터 없이 비디오만 입력받는 Content-Based Video Retrieval(CBVR) 연구에 해당한다. CBVR 은 비디오의 특징점을 추출하는 방법을 기준으로 크게 두 가지로 분류된다. 그중 fine-grained 접근법은 프레임 정보를 유지한 채로 비디오 사이의 기하학적 유사도를 계산한다. coarse-grained 접근법은 비디오의 유사도를 계산하기 위한 비디오 벡터로 비디오 레벨 정보를 활용한다. 프레임 정보를 종합하여 특징을 추출하고, 그 사이의 기하학적 유사도를 벡터의 내적과 같이 비교적 단순한 연산으로 계산한다. 비디오 사이의 연산이 상대적으로 단순하여 fine-grained 접근법에 비해 계산비용이 낮으나, 정확도가 낮다.

DnS[1]는 이미지 정보만 활용하는 비디오 탐색을 수행하며, fine 과 coarse-grained 접근법으로 얻은 유사도를 모두 활용한다. 유사도 사이의 차이가 작을 경우 계산효율을 위해 coarse-grained 유사도를 선택하고, 클 경우 fine-grained 유사도를 선택한다.

본 논문은 이미지 정보에서 얻는 유사도들의 차이가 너무 클 경우 오디오 정보 기반 유사도가 성능 개선에 기여할 수 있다는 것을 실험으로 보여준 연구[7]에 대한 후속연구이다. 본 논문에서는 이전 연구[7]에서 분석된 결과를 토대로 fine-grained, coarse-grained, 그리고 audio 정보 기반 유사도를 모두 활용해 최적의 유사도를 선택하는 Audio-visual switch network (SwitchNet)를 제안한다.

Ⅱ. 본론

제안하는 네트워크의 기반이 되는 DnS[1]에서는 비디오의 유효 이미지 정보를 CNN 기반 network 를 통해 추출하고, Knowledge distillation 을 활용해 효율적으로 Coarse-grained similarity(s^c)와 Fine-grained similarity(s^f)를 얻고, 얻은 유사도의 차이가 작을 때, 계산 비용이 적은 s^c를 선택하고, 클 때는 더 정확도가 높은 s^f를 선택한다.

본 논문에서는 오디오 정보를 추가로 활용하기 위해 오디오 기반 유사도(s^a)를 선택하는 기준을 추가하였다. 각 샘플에 대해 s^f, s^c, s^a 중 어떤 것을 최종 유사도로 활용할지 선택하는 라벨을 설정하며, 이 라벨은 아래 식과 같이 할당된다.

$$\begin{cases} label = 2, \ d(s^f, s^c) > threshold \\ label = 1, \ otherwise \\ label = 0, \ d(s^f, s^c) < \varepsilon \end{cases} \quad (1)$$

SwitchNet 에서는 비디오 입력에 CNN 을 통한 feature 추출, feature 사이의 tensor dot 연산, average pooling 을 수행하여, self-similarity matrix 를 얻고, 이를 CNN 에 전달한다. 최종 average pooling layer 를 통해 비디오 pair 각각에 대한 self-similarity score 를 얻고, 미리 계산해둔 coarse-

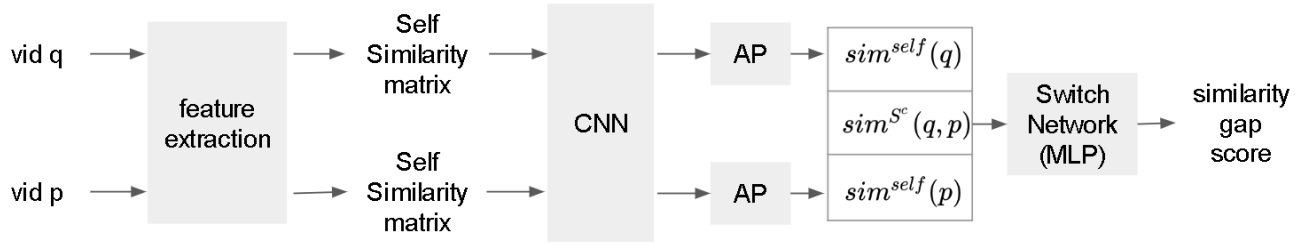

그림 1. SwitchNet 의 전반적인 구조

grained 유사도와 함께 MLP 구조의 SwitchNet 으로 전달하여 similarity gap score 를 예측하는 분류를 수행한다. 이때, 손실함수로 similarity gap score 와 (1)의 label 사이의 cross entropy 를 활용한다. 이를 통해 coarse, fine, audio 기반 유사도 중 어떤 유사도를 선택할지 학습하게 된다. 그림 1 은 해당 framework 를 나타낸다.

SwitchNet 의 성능 검증을 위해 설정한 세부 내용들은 다음과 같다. 먼저 시각적 정보 기반 유사도를 계산하기 위해, VCDB[3]를 triplet 손실함수로 학습한 ViSiL[4]을 teacher 로 두었다. 이후, DnS100K[1]에 대한 유사도를 teacher 에서 얻고 response-based knowledge distillation 으로 활용하여, fine-grained 와 coarse-grained student network 를 학습하였고, 해당 network 로 각 기법에 대한 시각적 정보 기반 유사도를 계산하였다. Audio 정보 기반 유사도는 VCDB 의 오디오로 학습된 AuSiL[2]을 활용한다.

	SVD(mAP)
TMK[5]	0.774
LAMV[6]	0.786
DnS[1]	0.902
SwitchNet(0.99)	0.910
SwitchNet(0.95)	0.892
SwitchNet(0.90)	0.867

표 1. SVD 에 대한 mean average precision 비교

SwitchNet 의 성능은 SVD[5]로 검증하였다. 표 1 은 SwitchNet 과 다른 최근 모델들의 SVD 에서의 성능을 비교한다. SwitchNet 은 여러 audio threshold 에 대해서 추가 실험하였다. Threshold 를 낮춰서, audio 기반 유사도가 선택될 기준을 완화할 수록 성능이 낮으며, 이는 이미지 기준의 annotation 에 대해 audio 기반 유사도를 선택하여, Audio-Visual Correspondence (AVC) noise 가 개입될 여지를 높이기 때문이다.

Ⅲ. 결론

본 논문에서는 knowledge distillation 을 통해 시각적 정보 기반 유사도를 효율적으로 계산하고, 해당 시각적 정보에 대한 신뢰도를 학습하여, 각 시각적 기법으로 얻은 유사도 사이의 차이가 너무 클 때, 오디오 정보를 선택하도록 학습한 SwitchNet 을 통해 시각적 정보와 오디오 정보를 적절히 선택하는 multi-modal video retrieval network 를 제안한다.

AVC noise 로 인해 audio 정보와 visual 정보의 high-level feature 가 상이한 경우가 빈번하며, 이는 audio-visual task 의 main challenge 이다. Audio-visual correspondence 를 오디오 선택 기준을 학습하는 데 활용하는 것으로 audio 활용 폭을 높이는 것을 향후 연구 방향으로 제안한다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(NO. 2021-0532, 다목적 비디오 검색을 위한 차세대 인공신경망 기술 개발)과 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-01341, 인공지능대학원지원(중앙대학교))

참 고 문 헌

[1] Kordopatis-Zilos, Giorgos, et al. "DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval."International Journal of Computer Vision 130.10 (2022): 2385-2407.

[2] Avgoustinakis, Pavlos, et al. "Audio-based near-duplicate video retrieval with audio similarity learning" 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.

[3] Jiang, Yu-Gang, Yudong Jiang, and Jiajun Wang. "VCDB: a large-scale database for partial copy detection in videos." European conference on computer vision. Springer, Cham, 2014.

[4] Kordopatis-Zilos, Giorgos, et al. "ViSiL: Fine-grained spatio-temporal video similarity learning." IEEE/CVF International Conference on Computer VIsion. 2019.

[5] Poullot, Sebastien, et al. "Temporal matching kernel with explicit feature maps." 23rd ACM international conference on Multimedia. 2015.

[6] Baraldi, Lorenzo, et al. "LAMV: Learning to align and match videos with kernelized temporal layers." IEEE conference on computer vision and pattern recognition. 2018.

[7] Jongsu Youn, Soohyun Park, Jongwook Choi, Jongwon Choi. "Deep Learning-based Video Retrieval with Audio-video Switching Network." IEIE Summer Annual Conference, 2022.