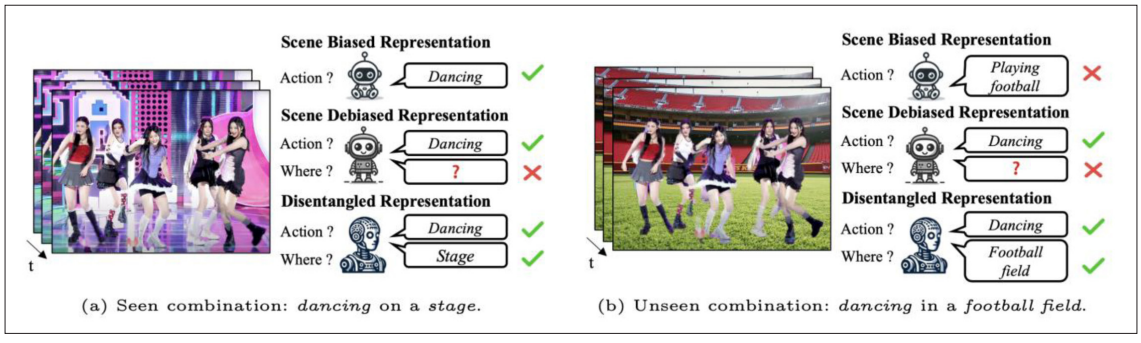
졸업논문 소개

액션과 장면의 분리된 비디오 표현 학습

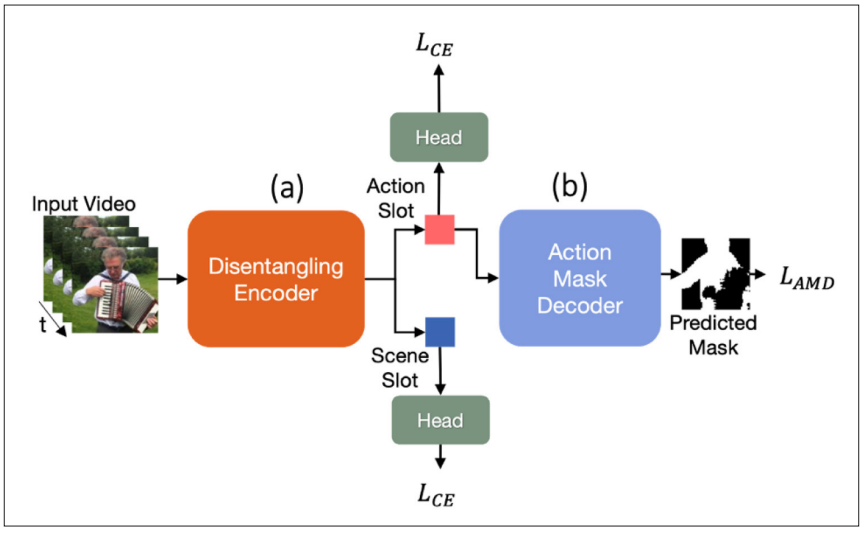
배경호 / 경희대학교 Vision and Learning Lab.

인간은 주변의 장면 컨텍스트에서 인간의 행동을 추출함으로써 자연스럽게 비디오의 내용을 이해할 수 있다. 이전에 보지 못한 행동-장면 조합을 만나더라도, 쉽게 비디오를 인식할 수 있다. 예를 들어, 〈그림 1〉(b)에서 사람들이 축구장에서 춤추는 모습을 볼 수 있다. 인간은 이러한 비디오를 만나도 행동과 장면을 쉽게 분리하여 이해할 수 있지만, 대부분의 비디오 행동 인식 모델들은 입력 비디오에서 행동과 장면 컨텍스트를 분해하는 데 어려움을 겪는다. 오히려, 비디오 행동 인식 모델들은 비디오 데이터셋에서 행동과 장면 사이의 허위 상관관계로 인해 장면에 편향된 행동 표현을 학습하는 경향이 보인다. 가장 큰 이유는 기존의 비디오 데이터셋들은 각 행동 클래스마다 제한된 행동-장면 쌍 조합을 주로 포함하고 있으며, 이는 현실에서 다양한 행동이 축구장에서도 발생할 수 있음에도 불구하고, 데이터셋 구축의 높은 비용 때문에 다양한 조합을 포함하지 못하기 때문이다.

이전의 연구는 장면 컨텍스트를 의도적으로 무시하는 장면-편향 제거 학습법(scene debiased representation)



<그림 1> Motivation



<그림 2> Overview

이 주를 이루었다. 장면에 대한 편향을 제거하면 〈그림 1〉(b)와 같은 상황에서 춤추는 행동은 올바르게 예측할 수 있지만 장면이 주는 풍부한 컨텍스트를 사용하지 못한다는 한계점이 있다.

이에 따라, 우리는 이전 연구들의 한계인 장면 컨텍스트를 무시하는 것이 아니라, 행동과 장면의 분리된 표현 학습(Disentangled VIdeo representations of Action and Scene)을 통한 종합적인 비디오 이해 문제에 접근한다. 이는 장면-편향된 행동 인식 모델이 미처 보지 못한 행동-장면 조합을 만났을 때 잘못된 예측을 할 가능성을 줄이며, 동시에 장면 정보를 고려함으로써 보다 풍부한 정보를 제공한다. 예를 들어, 분리된 표현을 통해 모델은 사람들이 무대(〈그림 1〉(a))나 축구장(〈그림 1〉(b))에서 춤추고 있다는 것을 정확하게 인식할 수 있다.

행동과 장면을 분리하여 학습하기 위해 우리는 DEVIAS라는 새로운 인코더-디코더 아키텍처를 제안한다. DEVIAS는 Disentangling Encoder(〈그림 2〉(a)), Action Mask Decoder(〈그림 2〉(b)), 그리고 액션/장면 분류 헤드로 구성되어 있다. 행동과 장면을 분리하여 학습하기 위한 핵심은 훈련 시간 동안 DE와 AMD를 모두 사

<표 1> UCF-101 dataset에서의 분리된 표현 측정 실험

Training Strategy	Supervision		Method	Action		Scene		H.M.
	Action	Scene		Seen ①	Unseen ②	Seen ③	Unseen ④	
Single-Task	✓	X	Action ViT	92.9	16.7	-	-	-
	✓	X	Scene-debiased	92.3	25.4	-	-	-
	X	✓	Scene ViT	-	-	72.0	64.7	-
Multi-Task	✓	✓	One-Token	91.9	13.4	74.0	61.7	34.7
	✓	✓	Two-Token	86.0	15.5	72.3	62.0	37.7
Disentangle	✓	✓	DEVIAS	90.1	39.4	74.0	64.5	**61.1**

<표 2> 다양한 downstream task 실험

Pretraining Strategy	Method	Temporal-biased		Scene-biased		H.M.
		Diving48	SSV2	UCF-101	ActivityNet	
Single-Task	Naive Action ViT	81.5	74.2	98.5	84.4	83.8
	BE	81.9	74.5	98.3	84.6	84.0
	FAME	80.6	74.2	98.3	83.8	83.4
	Naive Scene ViT	73.1	71.8	92.0	73.1	76.7
Multi-Task	Two-Token	80.1	73.7	98.2	83.7	83.0
	Two-Token w/ FAME	78.7	73.5	98.1	81.5	82.0
Disentangle	DEVIAS	84.4	75.2	98.4	84.5	**84.8**

용하는 것이다. 슬롯 주의를 사용하여 분리된 액션 및 장면 표현을 학습하고, AMD는 액션 슬롯을 입력으로 받아 액션 마스크를 예측하는 가벼운 디코더이다.

DEVIAS의 효과를 검증하기 위해 우리는 제어된 실험 세트를 신중하게 설계하였다.

먼저 〈표 1〉에서 행동과 장면을 올바르게 분리하였는지 검증하기 위해 행동과 장면 인식 성능을 모두 측정한다. 다른 비교 모델에 비해 DEVIAS는 행동과 장면을 올바르게 인식하며 second best model에 비해 23.4%의 조화 평균 성능이 향상한다. 특히, 학습 시에 보지 못한 조합에도 강인하게 인식하는 것을 알 수 있다. 이는 우리가 제안한 모델이 행동과 장면을 올바르게 분리했음을 검증한다.

또한, 〈표 2〉에서는 다양한 downstream task에서 우리가 제안한 DEVIAS가 높은 성능을 달성함을 보여준다. 비교 모델에 비해 가장 높은 조화 평균 성능을 달성했다. 이는 행동과 장면을 분리하여 학습하는 것이 기존의 방법론보다 더욱 풍부한 representation을 학습함을 알 수 있다.

배 경 호

- *2022년 2월 : 경희대학교 컴퓨터공학과 학사*
- *2024년 2월 : 경희대학교 인공지능학과 석사*
- *2024년 3월 ~ : 경희대학교 Vision and Learning Lab. 연구원*
- *주관심분야 : Video understanding, Representation learning and Domain adaptation*