AVA: 시공간적으로 로컬라이즈된 원자 시각 액션의 비디오 데이터 세트

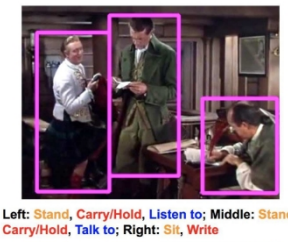
구춘희* 첸 선* 데이비드 A. 로스* 칼 본드릭* 캐롤라인 판토파루*
예칭 리* Sudheendra 비자야나라심한* 조지 토데리치* 수잔나 리코*

라훌 석탄카* 코델리아 슈미드 †* 지텐드라 말릭 ‡*

초록

이 백서에서는 시공간적으로 로컬라이즈된 원자 시각 액션(AVA)의 비디오 데이터세트를 소개합니다. AVA 데이터 세트는 15분 분량의 437개 비디오 클립에 80개의 원자적 시각적 동작을 조밀하게 주석으로 달고 있으며, 동작이 시공간적으로 로컬라이즈되어 있어 한 사람당 여러 개의 레이블이 자주 발생하는 159만 개의 동작 레이블이 생성됩니다. 데이터 세트의 주요 특징은 다음과 같습니다: (1) 복합적인 동작이 아닌 원자적인 시각적 동작의 정의, (2) 각 사람에 대해 여러 개의 주석이 포함된 정확한 시공간적 주석, (3) 15분 분량의 비디오 클립에 걸쳐 이러한 원자적 동작에 대한 철저한 주석, (4) 연속적인 세그먼트에 걸쳐 시간적으로 연결된 사람들, (5) 동영상을 사용하여 다양한 동작 표현을 수집한다는 점입니다. 이는 일반적으로 짧은 비디오 클립의 복합 액션에 대해 드문드문 주석을 제공하는 시공간적 액션 인식을 위한 기존 데이터 세트에서 벗어난 것입니다.

사실적인 장면과 액션의 복잡성으로 인해 AVA는 액션 인식의 본질적인 어려움을 안고 있습니다. 이를 벤치마킹하기 위해 유니티는 현재의 최신 기법을 기반으로 액션 로컬라이제이션에 대한 새로운 접근 방식을 제시하고, JHMDB



니다
ㄷ 불
게 나
성을

유니티는 행동 인식 연구에 새로운 주석이 달린 비디오 데이터 세트인 AVA를 도입했습니다(그림 1 참조). 이 데이터는 1Hz의 샘플링 주파수에서 사람 중심으로 추출됩니다. 모든 사람은 바운딩 박스를 사용하여 위치가 지정되며, 부착된 레이블은 배우가 수행하는 (여러 가지) 동작에 해당합니다. 서기, 앉기, 걷기, 수영 등 배우의 포즈(주황색 텍스트)에 해당하는 동작이 하나 있고, 물체와의 상호작용(빨간색 텍스트) 또는 상호 작용에 대응하는 추가 동작이 있을 수 있습니다.

*Google 리서치
†인리아, 장 쿤츠만 연구소, 프랑스 그르노블
‡캘리포니아 버클리 대학교, 미국

그림 1. AVA 데이터 세트의 샘플 프레임에 있는 바운딩 박스 및 액션 주석. 각 바운딩 박스는 1개의 포즈 동작(주황색), 객체와의 0~3개의 상호작용(빨간색), 다른 사람과의 0~3개의 상호작용(파란색)과 연관되어 있습니다. 이러한 동작 중 일부는 정확한 레이블을 지정하기 위해 시간적 컨텍스트가 필요합니다.

다른 사람과의 동작(파란색 텍스트). 여러 배우가 포함된 프레임의 각 인물은 개별적으로 레이블이 지정됩니다.

사람이 수행한 행동에 라벨을 붙이기 위해 중요한 선택은 행동이 분류되는 시간적 세분성에 따라 결정되는 어노테이션 어휘입니다. 저희는 짧은 세그먼트(키프레임을 중심으로 ±1.5초)를 사용하여 중간 프레임에서 행동에 라벨을 지정하기 위한 시간적 컨텍스트를 제공합니다. 이를 통해 주석 작성자는 정적 프레임에서 확인할 수 없는 집기 또는 내려놓기와 같은 동작을 명확히 구분하기 위해 움직임 단서를 사용할 수 있습니다. 시간적 컨텍스트를 비교적 짧게 유지하는 이유는 물리적 동작에 대한 (시간적으로) 세밀한 주석에 관심이 있기 때문이며, 이는 '원자적 시각적 동작(AVA)'의 동기가 됩니다. 음성 데이터는 80개의 서로 다른 원자적 시각 액션으로 구성됩니다. 데이터 세트는 437개의 서로 다른 영화에서 15분에서 30분 간격으로 추출한 것으로, 1Hz 샘플링 주파수를 고려하면 각 영화에 대해 900개의 키프레임을 제공합니다. 각 키프레임에서 모든 인물은 AVA 어휘의 동작(여러 개일 수도 있음)으로 레이블이 지정됩니다. 각 인물은 연속된 키프레임에 연결되어 동작 레이블의 짧은 시간적 시퀀스를 제공합니다(섹션 4.3). 이제 우리는

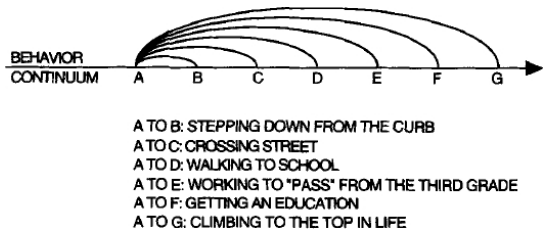


그림 2. 이 그림은 활동의 계층적 특성을 보여줍니다. 바커와 라이트 [3], 247페이지에서 발췌.

AVA의 주요 디자인 선택 사항입니다.

원자적 행동 범주. 바커와 라이트[3]는 캔자스주의 한 작은 마을 주민들의 일상 생활에서 나타나는 '행동 에피소드'에 대한 고전적인 연구에서 활동의 계층적 특성에 주목했습니다(그림 2). 가장 세밀한 수준에서 행동은 원자적인 신체 움직임이나 물체 조작으로 구성되지만, 더 거친 수준에서는 의도성과 목표 지향적 행동이라는 측면에서 가장 자연스럽게 설명할 수 있습니다.

이러한 계층 구조는 행동 어휘를 정의하기 어렵게 만들고, 물체 인식에 비해 이 분야의 발전이 더딘 이유이기도 하며, 높은 수준의 행동 에피소드를 일일이 나열하는 것은 비실용적입니다. 그러나 미세한 시간 척도로 제한하면 동작은 본질적으로 매우 물리적이며 시각적 징후가 명확합니다. 여기서는 키프레임에 1Hz로 주석을 달았는데, 이는 동작의 완전한 의미적 내용을 포착하기에 충분히 밀도가 높으면서도 동작 경계의 비현실적으로 정확한 시간적 주석을 요구하지 않을 수 있기 때문입니다. THUMOS 과제[18]에서는 객체와 달리 동작 경계는 본질적으로 모호하기 때문에 주석자 간에 상당한 의견 불일치가 발생한다는 것을 관찰했습니다. 이와 대조적으로, 주석자는 프레임에 특정 동작이 *포함되어* 있는지 여부를 ±1.5초의 컨텍스트를 사용하여 쉽게 판단할 수 있습니다. AVA는 사람 중심의 동작 시계열에서 허용 가능한 *정밀도인* ±0.5*초로* 동작 시작점과 종료점을 효과적으로 로컬라이즈합니다. 나무가 쓰러지는 것과 같은 이벤트는 사람이 관여하지 않지만, 우리는 단일 행위자로 취급되는 사람의 행동에 초점을 맞춥니다. 스포츠에서처럼 여러 사람이 있을 수도 있고 두 사람이 포옹할 수도 있지만, 각 사람은 개별적인 선택권을 가진 에이전트이기 때문에 각 세그먼트를 개별적으로 취급합니다. 시간에 따라 사람에게 할당된 행동 레이블은 시간적 모델링을 위한 풍부한 데이터 소스입니다(섹션 4.3).

동영상 주석. 이상적으로는 '야생'에서의 행동을 원할 것입니다. 하지만 영화는 특히 장르의 다양성과 영화 산업이 번성하는 국가를 고려할 때 매력적인 근사치입니다. 이 과정에서 약간의 편견이 있을 수 있습니다. 스토리는 흥미로워야 하며, 샷의 병치를 통해 소통하는 영화 언어의 문법[2]이 있습니다. 즉, 각 장면에서 우리는 유능한 배우가 연기하는 현실을 어느 정도 대변하는 일련의 인간 행동을 기대할 수 있습니다. AVA는 사용자가 생성한 비디오에서 소싱한 현재 데이터 세트를 보완합니다.

다양한 스토리텔링에 걸맞게 영화에 더 많은 활동을 담을 수 있습니다.

철저한 행동 라벨링. 모든 키프레임에 모든 인물의 모든 행동에 라벨을 붙입니다. 이렇게 하면 자연스럽게 행동 범주 간에 집프의 법칙에 따른 불균형이 다시 발생합니다. 기억에 남는 동작(춤추는 동작)보다 일반적인 동작(서 있거나 앉아있는 동작)의 예가 훨씬 더 많을 수 있지만, 이렇게 해야 합니다! 인식 모델은 인위적으로 균형 잡힌 데이터 세트를 사용하여 스캐폴딩하는 것이 아니라 현실적인 "긴 꼬리" 행동 분포[15]에서 작동해야 합니다. 프로토콜의 또 다른 결과는 인터넷 동영상 리소스를 명시적으로 쿼리하여 동작 범주의 예를 검색하지 않기 때문에 특정 종류의 편견을 피할 수 있다는 것입니다. 예를 들어, 문을 여는 동작은 영화 클립에서 자주 발생하는 일반적인 이벤트이지만 YouTube에서 해당 태그가 지정된 문 여는 동작은 비정형적이라는 점에서 주목할 만한 가치가 있습니다.

사실적인 복잡성을 지닌 AVA는 이 분야에서 널리 사용되는 많은 데이터 세트에 숨겨진 행동 인식의 내재적 어려움을 해결한다고 생각합니다. 예를 들어, 일반적인 배경에서 수영과 같이 시각적으로 눈에 띄는 동작을 하는 사람의 비디오 클립은 달리기하는 사람과 쉽게 구별할 수 있습니다. 이미지 크기가 작은 여러 명의 배우가 물체를 만지거나 잡는 등 미묘하게 다른 동작을 수행하는 AVA와 비교해보십시오. 이러한 직관을 검증하기 위해 JHMDB [20], UCF101-24 카테고리 [32] 및 AVA에 대한 비교 벤치마킹을 수행합니다. 시공간적 동작 로컬라이제이션에 사용하는 접근 방식(섹션 5 참조)은 멀티프레임 접근 방식 [16, 41]을 기반으로 하지만, I3D 컨볼루션[6]으로 튜블러를 분류합니다. JHMDB [20] 및 UCF101-24 카테고리 [32](섹션 6 참조)에서 최첨단 성능을 얻는 반면 AVA의 mAP는 15.8%에 불과합니다.

AVA 데이터 세트는 https에서 공개적으로 공개되었습니다: //research.google.com/ava/.

2. 관련 작업

동작 인식 데이터셋. KTH [35], Weizmann [4], Hollywood-2 [26], HMDB [24], UCF101 [39] 등 가장 널리 사용되는 액션 분류 데이터 세트는 하나의 액션을 캡처하기 위해 수동으로 트리밍된 짧은 클립으로 구성되어 있습니다. 이러한 데이터 세트는 완전 감독, 전체 클립, 강제 선택 비디오 분류기를 훈련하는 데 이상적입니다. 다시 말해, TrecVid MED [29], Sports- 1M [21], YouTube-8M [1], Something-something [12], SLAC [48], Moments in Time [28], Kinetics [22] 등의 데이터 세트는 대규모 비디오 분류에 중점을 두고 있으며, 종종 자동적으로 생성된 - 따라서 노이즈가 있을 수 있는 - 주석이 포함되어 있습니다. 이러한 기술들은 가치 있는 목적을 제공하지만 AVA와는 다른 요구 사항을 해결합니다.

최근의 일부 작업은 시간적 현지화를 향해 나아가고 있습니다. ActivityNet [5], THUMOS [18], MultiTHUMOS [46], Charades [37]는 트리밍되지 않은 대량의 비디오를 사용합니다,

각 데이터 세트에는 YouTube(ActivityNet, THUMOS, MultiTHUMOS) 또는 크라우드소싱된 액터(Charades)로부터 얻은 여러 액션이 포함되어 있습니다. 이 데이터 세트는 관심 있는 각 액션에 대한 시간적(공간적) 로컬라이제이션을 보여줍니다. AVA는 동작을 수행하는 각 피사체에 대해 시공간적 주석을 제공하고 주석이 15분 클립에 걸쳐 밀집되어 있다는 점에서 이들과 다릅니다.

CMU [23], MSR Actions [47], UCF Sports [32], JHMDB [20]와 같은 몇몇 데이터 세트는 짧은 동영상에 대해 각 프레임에 시공간적 주석을 제공합니다. AVA 데이터 세트와의 주요 차이점은 동작 수가 적 다 는 점 , 비디오 클립 수가 적다는 점, 클립이 매우 짧다는 점입니다. 또한 동작은 AVA에서처럼 원자 단위가 아닌 복합적(예: 장대높이뛰기)으로 구성됩니다. 최근 확장된 UCF101 [39], DALY [44], Hol- lywood2Tubes [27]는 트림되지 않은 비디오에서 시공간적 로컬라이제이션을 평가하기 때문에 작업이 상당히 어려워지고 성능이 저하될 수 있습니다. 그러나 동작 어휘는 여전히 제한된 수의 복합 동작으로 제한됩니다. 또한 동작을 촘촘하게 다루지 못하기 때문에 덩크슛을 하는 선수에게만 주석이 달린 UCF101의 BasketballDunk가 좋은 예입니다. 그러나 실제 애플리케이션에서는 모든 사람의 원자적인 행동에 대한 지속적인 주석이 필요한 경우가 많으며, 이를 상위 수준의 이벤트로 구성할 수 있습니다. 이것이 바로 AVA가 15분 클립에 걸쳐 철저한 라벨링을 하는 이유입니다.

AVA는 두 가지 측면에서 제한이 있는 정지 이미지 동작 인식 데이터 세트[7, 9, 13]와도 관련이 있습니다. 첫째, 움직임이 없기 때문에 동작을 명확하게 구분하기 어렵습니다. 둘째, 정지 이미지에서는 복합 이벤트를 원자 동작의 *시퀀스로* 모델링하는 것이 불가능합니다. 이는 여기서는 다루지 않을 수도 있지만, AVA가 학습 데이터를 제공하는 많은 실제 애플리케이션에서는 분명히 필요합니다.

시공간적 동작 로컬라이제이션 방법. 가장 최근의 접근 방식[11, 30, 34, 43]은 2스트림 변형으로 프레임 수준에서 동작 클래스를 구분하도록 훈련된 오브젝트 감지기를 사용하여 RGB 및 흐름 데이터를 순차적으로 처리합니다. 그런 다음 동적 프로그래밍[11, 38] 또는 추적[43]을 사용하여 프레임별 감지 결과를 연결합니다. 이러한 모든 접근 방식은 프레임 수준 감지를 통합하는 데 의존합니다. 최근에는 다중 프레임 접근 방식이 등장했습니다: 튜브[41]는 여러 프레임에 걸쳐 로컬라이제이션과 분류를 공동으로 추정하고, T-CNN[16]은 3D 컨볼루션을 사용하여 짧은 튜브를 추정하며 , 마이크로 튜브는 두 개의 연속된 프레임에 의존하고[33], 포즈 가이드 3D 컨볼루션은 두 개의 스트림 접근 방식에 포즈를 추가합니다[49]. 유니티는 시공간 튜브의 아이디어를 기반으로 하지만, 최첨단 I3D 컨볼루션[6] 및 더 빠른 R-CNN[31] 영역 제안을 사용하여 최첨단 기술을 능가하는 성능을 발휘합니다.

3. 데이터 수집

AVA 데이터 세트의 주석은 액션 어휘 생성, 동영상 및 세그먼트 선택의 5단계로 구성됩니다,

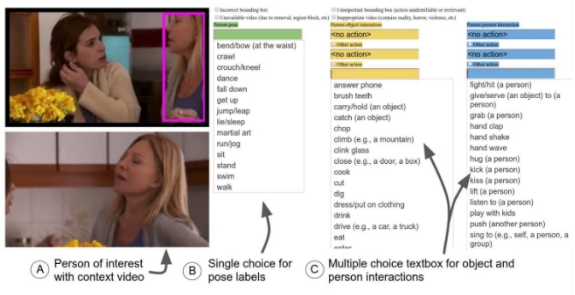


그림 3. 작업 주석을 위한 사용자 인터페이스. 자세한 내용은 3.5 절에 나와 있습니다.

사람 경계 상자 주석, 사람 연결 및 작업 주석을 추가할 수 있습니다.

3.1. 액션 어휘 생성

트위터는 액션 어휘를 생성하기 위해 세 가지 원칙을 따릅니다. 첫 번째는 일반성입니다. 특정 환경에서의 특정 활동(예: 농구 코트에서 농구하기)이 아닌 일상 생활 장면에서 일반적인 행동을 수집합니다. 두 번째는 원자성입니다. 액션 클래스는 명확한 시각적 시그니처를 가지고 있으며, 일반적으로 특정 객체와 독립적입니다(예: 어떤 객체를 잡을지 지정하지 않고 잡기). 따라서 목록이 짧으면서도 완전합니다. 마지막은 완전성입니다. 이전 데이터 세트의 지식을 사용하여 목록을 초기화하고, 어노테이터가 레이블을 지정한 AVA 데이터 세트의 동작 중 약 99%를 포함할 때까지 여러 차례 반복했습니다. 그 결과 어휘에 14개의 포즈 클래스, 49개의 사람-사물 상호작용 클래스, 17개의 사람-사람 상호작용 클래스가 포함되었습니다.

3.2. 동영상 및 세그먼트 선택

AVA 데이터 세트의 원시 비디오 콘텐츠는 YouTube에서 가져옵니다. 먼저 다양한 국적의 유명 배우 목록을 수집합니다. 각 이름에 대해 YouTube 검색 쿼리를 실행하여 최대 2000개의 결과를 검색합니다. '영화' 또는 '텔레비전' 주제 주석이 있고, 동영상이 30분 이상이며, 업로드된 지 1년 이상이고, 조회수가 1000회 이상인 동영상만 포함합니다. 또한 흑백, 저해상도, 애니메이션, 만화, 게임 동영상 및 성인용 콘텐츠가 포함된 동영상은 제외됩니다.

제약 조건 내에서 대표적인 데이터 세트를 만들기 위해, 트위터의 선택 기준은 액션 키워드로 필터링하거나 자동화된 액션 분류기를 사용하거나 일률적인 레이블을 강제 적용하는 것을 피합니다. 저희는 대형 영화 산업에서 샘플링하여 국제적인 영화 컬렉션을 만드는 것을 목표로 합니다. 그러나 영화 속 액션 묘사는 성별 등에 따라 편향되어 있으며[10], 인간 활동의 '실제' 분포를 반영하지 못합니다.

15분에서 30분 사이의 하위 부분에만 레이블을 지정하므로 각 영화는 데이터 세트에 동일하게 기여합니다. 제목이나 예고편에 주석을 달지 않기 위해 영화의 시작 부분을 건너뜁니다. 15분 길이를 선택하면 정해진 주석 예산으로 더 많은 영화를 포함할 수 있으므로 데이터 세트의 다양성을 높일 수 있습니다.

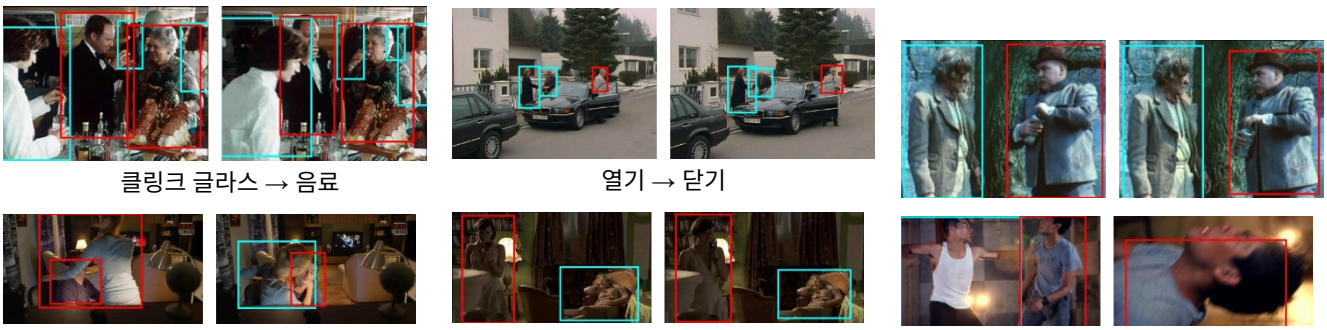


그림 4. AVA에서 원자 행동이 시간에 따라 어떻게 변하는지를 보여주는 예입니다. 텍스트는 빨간색으로 표시된 사람들의 원자 행동 쌍을 보여줍니다.

경계 상자. 시간 정보는 많은 동작을 인식하는 데 핵심적인 요소이며, 문이나 병을 여는 것과 같이 동작 범주 내에서도 모양이 크게 달라질 수 있습니다.

그런 다음 15분 분량의 각 클립을 1초 간격의 3초짜리 동영상 세그먼트 900개로 분할합니다.

3.3. 사람 경계 상자 주석

바인딩 박스를 사용하여 인물과 그 사람의 행동을 현지화합니다. 키프레임에 여러 피사체가 있는 경우 각 피사체는 액션 주석을 위해 주석 작성자에게 개별적으로 표시되므로 각 피사체의 액션 레이블이 다를 수 있습니다.

바운딩 박스 주석은 수작업이 많이 필요하기 때문에 하이브리드 접근 방식을 선택합니다. 먼저, Faster-RCNN 사람 검출기[31]를 사용하여 초기 경계 상자 집합을 생성합니다. 높은 정밀도를 보장하기 위해 작동점을 설정합니다. 그런 다음 어노테이터가 감지기가 놓친 나머지 바운딩 박스에 주석을 달면 됩니다. 이 하이브리드 접근 방식은 벤치마킹에 필수적인 전체 바운딩 박스 리콜을 보장하는 동시에 수동 주석에 드는 비용을 최소화합니다. 이 수동 주석은 사람 감지기가 놓친 바운딩 박스를 5%만 더 검색하여 디자인 선택의 유효성을 검증합니다. 잘못된 바운딩 박스는 다음 단계의 액션 어노테이션에서 어노테이터가 표시하고 다시 이동합니다.

3.4. 사람 링크 주석

바운딩 박스를 단기간에 걸쳐 연결하여 지상 실측 인물 트래클릿을 얻습니다. 사람 임베딩[45]을 사용하여 인접한 키 프레임의 바운딩 박스 간의 쌍별 유사도를 계산하고 헝가리 알고리즘[25]을 사용하여 최적의 매칭을 찾습니다. 자동 매칭은 일반적으로 강력하지만, 우리는 각 매칭을 검증하는 사람 주석자를 통해 오탐을 추가로 제거합니다. 이 절차를 통해 몇 초에서 몇 분 사이에 81,000개의 트랙릿이 호출됩니다.

3.5. 액션 어노테이션

액션 레이블은 그림 3에 표시된 인터페이스를 사용하여 크라우드 소싱된 애니메이터가 생성합니다. 왼쪽 패널에는 대상 세그먼트의 중간 프레임(위)과 반복되는 임베디드 비디오(아래)로서의 세그먼트가 모두 표시됩니다. 중간 프레임에 오버레이된 경계 상자는 레이블을 지정해야 하는 동작을 수행하는 사람을 지정합니다. 오른쪽

는 포즈 동작 1개(필수), 사람-사물 상호작용 3개(선택), 사람-사람 상호작용 3개(선택)를 포 함 하 여 최대 7개의 동작 레이블을 입력할 수 있는 텍스트 상자입니다. 나열된 동작 중 설명할 수 있는 동작이 없는 경우, 주석 작성자는 '기타 동작'이라는 확인란에 플래그를 지정할 수 있습니다. 또한 차단되었거나 부적절한 콘텐츠가 포함된 세그먼트 또는 잘못된 경계 상자에 플래그를 지정할 수 있습니다.

실제로 80개 클래스로 구성된 대규모 어휘에서 올바른 액션을 모두 찾으라고 지시하면 어노터가 올바른 액션을 놓치는 것이 불가피하다는 것을 관찰했습니다. 36]에서 영감을 받아 액션 주석 파이프라인을 액션 제안과 검증의 두 단계로 나누었습니다. 먼저 여러 명의 주석가에게 각 질문에 대한 액션 후보를 제안하도록 요청하여 공동 세트가 개별 제안보다 더 높은 리콜을 갖도록 합니다. 그런 다음 주석가들은 두 번째 단계에서 이러한 제안된 후보를 검증합니다. 이 2단계 접근 방식을 사용한 결과, 특히 예시가 적은 작업에 대해 상당한 리콜 개선 효과가 나타났습니다. 보충 자료에서 자세한 분석을 참조하세요. 평균적으로 어노테이터는 제안 단계에서 주어진 비디오 세그먼트에 주석을 다는 데 22초, 검증 단계에서는 19.7초가 소요됩니다
.

각 동영상 클립에는 세 명의 독립적인 주석가가 주석을 달며, 트위터에서는 최소 두 명의 주석가가 확인한 작업 라벨에 대해서만 근거 자료로 간주합니다. 주석 작성자에게는 무작위 순서로 세그먼트가 표시됩니다.

3.6. 교육, 검증 및 테스트 세트

교육/검증/테스트 세트는 비디오 수준에서 분할되므로 한 비디오의 모든 세그먼트가 한 분할에만 나타납니다. 437개의 비디오는 교육용 239개, 검증용 64개, 테스트용 134개로 약 55:15:30 비율로 분할되어 교육용 21만 5천 개, 검증용 5만 7천 개, 테스트용 12만 개 세그먼트가 생성됩니다.

4. AVA 데이터 세트의 특징

먼저 시각적 예시를 통해 AVA 데이터 세트의 다양성과 난이도에 대한 직관을 구축합니다. 그런 다음 데이터 세트의 주석을 정량적으로 특성화합니다. 마지막으로 행동과 시간적 구조를 탐색합니다.

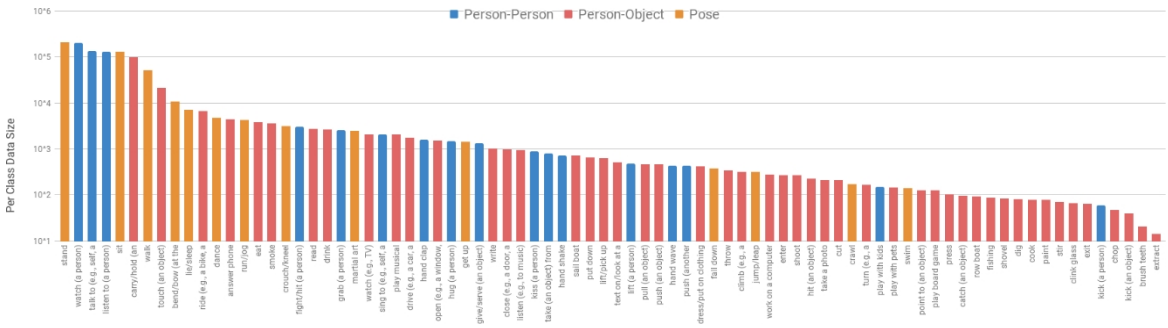


그림 5. 내림차순으로 정렬된 AVA 훈련/값 데이터 세트의 각 액션 클래스의 크기와 액션 유형을 나타내는 색상.

4.1. 다양성 및 난이도

그림 4는 연속된 세그먼트에 걸쳐 변화하는 원자 동작의 예를 보여줍니다. 바운딩 박스 크기와 시네마토그래피의 변화 외에도 '유리잔 부딪치기' 대 '마시기'와 같은 세밀한 차이를 구분하거나 '열기' 대 '닫기'와 같은 시간적 맥락을 활용해야 하는 경우가 많습니다.

그림 4는 "열기" 액션에 대한 두 가지 예시를 보여줍니다. 액션 클래스 내에서도 매우 다양한 컨텍스트에 따라 모양이 달라지며, 심지어 열리는 대상도 달라질 수 있습니다. 클래스 내 다양성을 통해 '열기'를 위한 봉인 해제와 같이 동작의 중요한 시공간적 부분을 식별하는 특징을 학습할 수 있습니다.

4.2. 주석 통계

그림 5는 AVA에서 액션 주석의 분포를 보여줍니다. 이 분포는 대략 지프의 법칙을 따릅니다. 그림 6은 바운딩 박스 크기 분포를 보여줍니다. 대부분의 사람들이 프레임의 전체 높이를 차지합니다. 그러나 여전히 작은 크기의 상자가 많이 있습니다. 이러한 가변성은 줌 레벨과 포즈 모두로 설명할 수 있습니다. 예를 들어, '입력'이라는 레이블이 있는 상자는 평균 너비가 이미지 너비의 30%, 평균 높이가 72%인 1:2의 일반적인 보행자 가로세로 비율을 보여줍니다. 반면에 '누워/수면'이라고 표시된 상자는 정사각형에 가깝고 평균 너비가 58%, 높이가 67%입니다. 상자 너비가 넓게 분포되어 있어 사람들이 라벨이 붙은 동작을 실행하기 위해 취하는 다양한 포즈를 보여줍니다.

대부분의 사람 바운딩 박스에는 여러 개의 레이블이 있습니다. 모든 경계 상자에는 하나의 포즈 레이블이 있고, 28%의 경계 상자에는 최소 1개의 사람-물체 간 동작 레이블이 있으며, 67%의 경계 상자에는 최소 1개의 사람-사람 간 상호 작용 레이블이 있습니다.

4.3. 시간 구조

AVA의 주요 특징은 세그먼트에서 세그먼트로 진화하는 풍부한 시간적 구조입니다. 세그먼트 간에 사람들을 연결했기 때문에 같은 사람이 수행한 행동 쌍을 살펴봄으로써 공통된 연속 행동을 발견할 수 있습니다. 정규화된 점순으로 쌍을 정렬합니다.

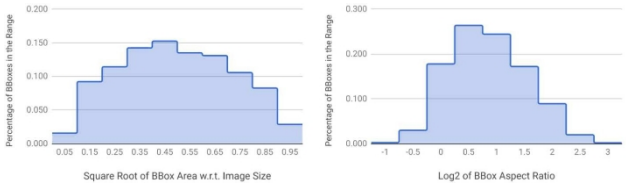


그림 6. AVA 데이터 세트에서 주석이 달린 바운딩 박스의 크기 및 종횡비 변화. 바운딩 박스는 다양한 크기로 구성되어 있으며, 그 중 상당수는 작아서 감지하기 어렵습니다. 바운딩 박스의 종횡비에도 큰 변화가 있으며, 2:1 비율의 모드(예: 앉은 자세)에서도 큰 변화가 나타납니다.

Mutual Information (NPMI) [8], which is commonly used in linguistics to represent the co-occurrence between two

단어: NPMI(x, y) = $\ln\frac{p(x,y)}{p(x)p(y)}$ ⁄ (- ln $p(x, y)$). 값은 직관적으로 [-1, 1] 범위에 속하며, 절대 함께 발생하지 않는 단어 쌍은 -1, 독립적인 단어 쌍은 0, 1이 됩니다.
를 사용하여 항상 함께 발생하는 쌍을 표시합니다.

표 1은 동일한 사람에 대해 연속적인 1초 세그먼트에서 NPMI가 가장 높은 작업 쌍을 보여줍니다. 신원 전환을 제거한 후 몇 가지 흥미로운 상식적인 시간적 패턴이 발생합니다. "전화 보기" → "전화 받기", "쓰러지다"로의 전환이 자주 발생합니다.
→ "거짓말" 또는 "듣다" → "말하다"로 바뀝니다. 또한 사람 간 행동 쌍도 분석합니다. 표 2는 동시에 형성되었지만 서로 다른 사람들에 의해 형성된 상위 행동 쌍을 보여줍니다. "타다" ↔ "운전하다", "음악 재생" ↔ "듣다", "가져가다" ↔ "주다/제공하다"와 같이 여러 의미 있는 쌍이 등장합니다. 상대적으로 거친 시간적 샘플링에도 불구하고 원자적 행동 사이의 전환은 더 긴 시간적 구조를 가진 더 복잡한 행동 및 활동 모델을 구축하는 데 훌륭한 데이터를 제공합니다.

5. 액션 현지화 모델

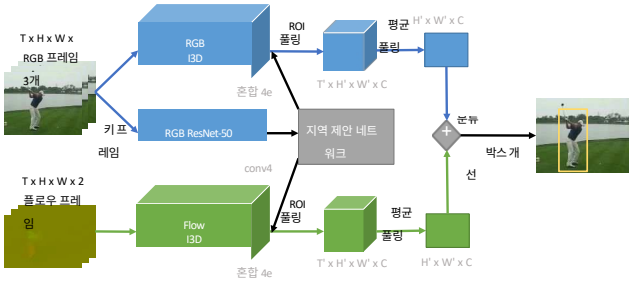
최근 몇 년 동안 UCF101이나 JHMDB와 같이 널리 사용되는 동작 인식 데이터 세트의 성능 수치가 상당히 높아졌지만, 이는 최첨단 기술에 대한 인위적인 장밋빛 그림을 제시할 수 있다고 생각합니다. 비디오 클립에서 한 사람만이 똑같이 특징적인 배경 장면에서 수영과 같이 시각적으로 특징적인 동작을 수행하는 경우, 행동을 분류하기는 쉽

첫 번째 조치	두 번째 작업 드라	NPMI
타기(예: 자전거/자동	이브(예: 자동차/트	0.68
차/말) 시청(예: TV)	럭)는 컴퓨터에서	0.64
운전(예: 자동차/트럭)	작동합니다.	0.63
열기(예: 창문/문)	타기(예: 자전거/자동차/말	0.59
휴대폰으로 문자 보내기/	) 닫기(예: 문/상자) 전화	0.53
보기 (사람) 듣기	받기	0.47
쓰러지다	(사람) 거짓말/수	0.46
(사람) 스탠드와 대	면과 대화	0.43
화	(사람) 앉아있음	0.40
걷기	듣기	0.40
	스탠드	

사람 1 행동	사람 2 동작 운전	NPMI
타기(예: 자전거/자동차	하다(예: 자동차/트	0.60
/말) 악기 연주하기 찍	럭) 듣다(예: 음악)	0.57
기(물건)	주다/섬기다(물건)	0.51
(사람) 스탠드와 대	듣다(사람) 앉다	0.46
화	댄스	0.31
악기 연주 산책	스탠드	0.23
(사람) 걷는 모	쓰기	0.21
습 보기	실행/	0.15
싸움/타격(사람)	조그	0.15
	스탠드	0.14

표 1. 다음과 같이 가능성이 높은 연속 작업의 상위 쌍을 보여줍니다.
같은 사람에 대해 이전/이후에 일어날 수 있습니다. NPMI를 기준으로 정렬합니다.

큐레이팅합니다. 배우가 여러 명일 때, 이미지 크기가 작을 때, 미묘하게 다른 동작을 수행할 때, 배경 장면이 무슨 일이 일어나고 있는지 알기에 충분하지 않을 때 어려움이 발생합니다. AVA에는 이러한 측면이 많이 있으며, 그 결과



AVA의 성능이 훨씬 떨어지는 것을 알 수 있습니다. 실제로 이 발견은 단어 맞추기 데이터 세트의 성능 저하로 인해 예고된 것이었습니다[37].

이를 증명하기 위해 멀티 프레임 시간 정보에서 작동하는 시공간적 동작 로컬라이제이션에 대한 최근 접근 방식에서 영감을 얻은 최첨단 동작 로컬라이제이션 접근 방식을 개발했습니다[16, 41]. 여기서는 동작 감지를 위해 I3D[6]에 기반한 더 큰 시간적 컨텍스트의 영향에 의존합니다. 접근 방식에 대한 개요는 그림 7을 참조하십시오.

Peng과 슈미드[30]에 이어, 저희는 엔드투엔드 로컬라이제이션과 동작 분류를 위해 더 빠른 RCNN 알고리즘[31]을 적용합니다. 그러나 이 접근 방식에서는 여러 프레임의 입력 채널이 시간이 지남에 따라 연결되는 첫 번째 레이어에서 시간 정보가 손실됩니다. 저희는 시간적 컨텍스트를 모델링하기 위해 Car- reira와 Zisserman[6]의 Inception 3D(I3D) 아키텍처를 사용할 것을 제안합니다. I3D 아키텍처는 Inception 아키텍처[40]를 기반으로 설계되었지만 2D 컨볼루션을 3D 컨볼루션으로 대체합니다. 시간적 정보는 네트워크 전체에 걸쳐 유지됩니다. I3D는 광범위한 비디오 분류 벤치마크에서 최첨단 성능을 달성합니다.

Faster RCNN과 함께 I3D를 사용하려면 모델을 다음과 같이 변경합니다. 먼저 I3D 모델에 길이 T의 입력 프레임을 공급하고 다음과 같은 3D 피처 맵을 추출합니다.

그림 7. 시공간적 동작 로컬라이제이션에 대한 접근 방식 그림. 영역 제안은 RGB 키프레임에서 Faster-RCNN으로 감지 및 회귀됩니다. 시공간 튜브는 2스트림 I3D 컨볼루션으로 분류됩니다.

표 2. 서로 다른 작업별로 동시 작업의 상위 쌍을 보여줍니다. 사람들. NPMI를 기준으로 정렬합니다.

넷 워크의 *혼합 4e* 레이어에서 $T' \times W' \times H' \times C$ 크기입니다. *혼합 4e의* 출력 피처 맵의 보폭은 16이며, 이는 ResNet [14]의 conv4 블록과 동일합니다. 둘째, 액션 제안 생성을 위해 키프레임에서 2D ResNet-50 모델을 영역 제안 네트워크의 입력으로 사용하여 입력 길이가 다른 I3D가 생성된 액션 제안의 품질에 미치는 영향을 피합니다. 마지막으로, 모든 시간 단계에 걸쳐 동일한 공간 위치에 2D ROI 풀링을 적용하여 ROI 풀링을 3D로 확장합니다. 액션 감지에 대한 광학적 흐름의 영향을 이해하기 위해 평균 풀링을 사용하여 피처 맵 수준에서 RGB 스트림과 광학적 흐름 스트림을 융합합니다.

기준선. AVA의 프레임 기반 2스트림 접근 방식과 비교하기 위해 [30]의 변형을 구현했습니다. ResNet-50[14]과 함께 Faster RCNN[31]을 사용하여 동작 제안과 액션 레이블을 공동으로 학습합니다. 영역 제안은 RGB 스트림으로만 유지됩니다. 영역 분류기는 5개의 연속 프레임에 걸쳐 쌓인 광학 흐름 특징과 함께 RGB를 입력으로 사용합니다. I3D 접근 방식에서는 conv4 특징 맵과 평균 풀링을 융합하여 RGB와 광학 흐름 스트림을 공동으로 훈련합니다.

구현 세부 사항. 광학적 흐름 특징을 추출하기 위해 FlowNet v2 [19]를 구현합니다. 비동기 SGD로 Faster-RCNN을 훈련합니다. 모든 훈련 작업에 대해 유효성 검사 집합을 사용하여 훈련 단계의 수를 결정하며, 그 범위는 600K에서 1백만 반복입니다. 입력 해상도는 320×400픽셀로 고정합니다. 다른 모든 모델 파라미터는 물체 감지를 위해 조정된 [17]의 권장 값을 기반으로 설정합니다. ResNet-50 네트워크는 ImageNet 사전 훈련된 모델로 초기화됩니다. 광학 흐름 스트림의 경우 conv1 필터를 복제하여 5프레임을 입력합니다. I3D 네트워크는 RGB 및 광학 흐름 스트림 모두에 대해 Kinetics [22] 사전 훈련된 모델로 초기화됩니다. I3D는 64프레임 입력으로 사전 학습되었지만 네트워크는 시간이 지남에 따라 완전히 컨볼 루션화되며 원하는 수의 프레임을 입력으로 사용할 수 있습니다. 모든 특징 레이어는 훈련 중에 공동으로 업데이트됩니다. 출력 프레임 수준 감지는 임계값 0.6의 비최대 억제를 사용하여 후처리됩니다.

AVA와 기존 액션 디텍션 데이터 세트의 주요 차이점 중 하나는 AVA의 액션 레이블이 뮤-

배타적입니다. 이 문제를 해결하기 위해 표준 소프트맥스 손실 함수를 각 클래스에 대해 하나씩 이진 시그모이드 손실의 합으로 대체합니다. AVA에는 시그모이드 손실을, 다른 모든 데이터 세트에는 소프트 최대 손실을 사용합니다.

연결. 프레임 단위로 탐지한 내용을 연결하여 액션 튜브를 구성합니다. 획득한 튜브에 대한 평균 점수를 기반으로 비디오 수준별 성능을 보고합니다. 시간적 라벨링을 적용하지 않는다는 점을 제외하고는 [38]에서 설명한 것과 동일한 연결 알고리즘을 사용합니다. AVA는 1Hz로 주석이 달리고 각 튜브에 여러 개의 라벨이 있을 수 있으므로 비디오 수준 평가 프로토콜을 수정하여 상한을 추정합니다. 기준 영상 링크를 사용하여 감지 링크를 추론하고, 기준 영상 튜브와 감지 튜브 사이의 클래스 IoU 점수를 계산할 때는 해당 클래스에 의해 레이블이 지정된 튜브 세그먼트만 고려합니다.

6. 실험 및 분석

이제 AVA의 주요 특성을 실험적으로 분석하고 행동 이해를 위한 도전 과제를 제시합니다.

6.1. 데이터 세트 및 메트릭

AVA 벤치마크. AVA의 레이블 분포는 대략 집프의 법칙(그림 5)을 따르고 매우 적은 수의 예제에 대한 평가는 신뢰할 수 없기 때문에 검증 및 테스트 분할에 최소 25개의 인스턴스가 있는 클래스를 사용하여 성능을 벤치마킹합니다. 결과 벤치마크는 60개 클래스에 대한 총 214,622개의 교육, 57,472개의 검증 및 120,332개의 테스트 예제로 구성됩니다. 별도의 언급이 없는 한, 훈련 세트에서 학습하고 검증 세트에서 평가한 결과를 보고합니다. 모델 파라미터 튜닝을 위해 훈련 데이터의 10%를 무작위로 선택합니다.

데이터 세트. AVA 외에도 난이도를 비교하기 위해 표준 비디오 데이터 세트도 분석합니다. JHMDB[20]는 21개 클래스에 걸쳐 928개의 트리밍된 클립으로 구성되어 있습니다. 저희는 절제 연구에서 분할 1에 대한 결과를 보고하지만, 최신 기술과의 비교를 위해 세 번의 분할에 대한 결과를 평균화했습니다. UCF101의 경우, Singh *외*[38]가 제공한 3207개의 비디오가 포함된 24개 클래스 하위 집합에 대한 시공간 주석을 사용합니다. 우리는 공식 분할1을 표준으로 실험을 수행합니다.

메트릭. 평가를 위해 가능한 경우 표준 관행을 따릅니다. 유니온은 프레임 수준과 비디오 수준에서 교차점 통과율(IoU)을 보고합니다. 프레임 수준 IoU의 경우, PASCAL VOC 챌린지[9]에서 사용하는 표준 프로토콜을 따르고 0.5의 IoU 임계값을 사용하여 평균 정밀도(AP)를 보고합니다. 각 클래스에 대해 평균 정밀도를 계산하고 모든 클래스에 대한 평균을 보고합니다. 비디오 수준 IoU의 경우, 임계값 0.5를 기준으로 기준점 튜브와 연결된 감지 튜브 간의 3D IoU를 계산합니다. 평균 AP는 모든 클래스에 대한 평균을 계산하여 계산합니다.

6.2. 최신 기술과의 비교

표 3은 두 가지 표준 비디오 데이터 세트에 대한 모델 성능을 보여줍니다. 3D 2스트림 모델은 상태 정보를 얻습니다
.

프레임 맵	JHMDB	UCF101-24
액션성 [42]	39.9%	-
펭, MR 제외 [30]	56.9%	64.8%
펭과 MR [30]	58.5%	65.7%
ACT [41]	65.7%	69.5%
접근 방식	73.3%	76.3%

비디오 맵	JHMDB	UCF101-24
펭과 MR [30]	73.1%	35.9%
Singh 외 [38]	72.0%	46.3%
ACT [41]	73.7%	51.4%
TCNN [16]	76.9%	-
접근 방식	78.6%	59.9%

표 3. 프레임 맵(상단) 및 비디오 맵(하단) @ IoU 0.5(JHMDB 및 UCF101-24). JHMDB의 경우, 세 개의 분할에 대한 평균 성능을 보고합니다. 퓨어스토리지의 접근 방식은 두 지표 모두에서 이전 최신 기술을 크게 능가합니다.

프레임 맵 및 비디오 맵 메트릭 모두에서 잘 정립된 기준선을 뛰어넘는 UCF101 및 JHMDB의 최첨단 성능을 제공합니다.

그러나 원자 동작을 인식할 때는 상황이 달라집니다. 표 4는 동일한 모델이 AVA 검증 세트(프레임 맵 15.8%, 비디오 맵 12.3%, 0.5 IoU에서 17.9%)와 테스트 세트(프레임 맵 14.7%)에서 상대적으로 낮은 성능을 얻는다는 것을 보여줍니다. 이는 AVA의 설계 원칙에 따른 것으로, 문맥과 객체 단서가 동작 인식에 변별력을 주지 않는 어휘를 수집했습니다. 대신 AVA에서 성공하려면 세분화된 세부 사항과 풍부한 시간적 모델을 인식해야 할 수 있으며, 이는 시각적 동작 인식에 새로운 도전 과제를 제시합니다. 이 백서의 나머지 부분에서는 AVA를 어렵게 만드는 요인을 분석하고 앞으로 나아갈 방향에 대해 논의합니다.

6.3. 절제 연구

AVA 카테고리를 인식하는 데 시간적 정보가 얼마나 중요할까요? 표 4는 시간적 길이와 모델 유형이 미치는 영향을 보여줍니다. 모든 3D 모델은 JHMDB 및 UCF101-24에서 2D 기준선보다 성능이 뛰어납니다. AVA의 경우 3D 모델은 10프레임 이상을 사용하면 더 나은 성능을 보입니다. 또한 템포럴 윈-다운의 길이를 늘리면 모든 데이터 세트에서 3D 투스트림 모델에 도움이 된다는 것을 알 수 있습니다. 예상대로 RGB와 광학 흐름 기능을 결합하면 단일 입력 방식에 비해 성능이 향상됩니다. 또한 AVA는 20프레임에서 성능이 포화 상태에 이르는 JHMDB 및 UCF101보다 더 큰 시간적 컨텍스트에서 더 많은 이점을 얻습니다. 이러한 이점과 표 1의 연속적인 작업은 AVA의 풍부한 시간적 컨텍스트를 활용하면 더 많은 이점을 얻을 수 있음을 시사합니다.

로컬라이제이션과 인식은 얼마나 어려울까요? 표 5는 엔드투엔드 액션 로컬라이제이션 및 인식과 클래스 불가지론적 액션 로컬라이제이션의 성능을 비교한 것입니다. 액션 로컬라이제이션이 더 어렵다는 것을 알 수 있습니다.

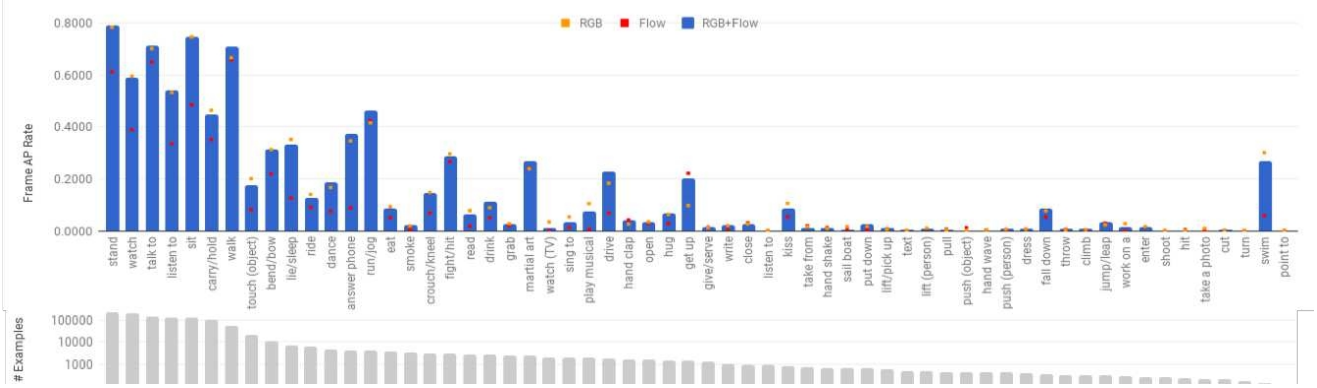


그림 8. 위: 각 작업 클래스에 대한 모델의 성능을 학습 예제 수에 따라 정렬하여 플롯합니다. 아래쪽: 클래스당 훈련 예제 수를 플롯합니다. 데이터가 많을수록 좋지만, 이상값은 모든 클래스의 복잡성이 동일하지는 않다는 것을 시사합니다. 예를 들어, 가장 작은 클래스 중 하나는 '수영'의 성능이 가장 높은 이유는 관련 장면의 상대적으로 쉽기 때문입니다.

프레임수	입력	JHMDB	UCF101-24	AVA
2D	1 RGB + 5 플로우	52.1%	60.1%	14.2%
3D	5 RGB + 5 플로우	67.9%	76.1%	13.6%
3D	10 RGB + 10 Flow	73.4%	78.0%	14.2%
3D	20 RGB + 20 Flow	76.4%	78.3%	14.8%
3D	40 RGB + 40 Flow	76.7%	76.0%	15.8%
3D	50 RGB + 50 Flow	-	73.2%	15.7%
3D	20 RGB	73.2%	77.0%	14.6%
3D	20 흐름	67.0%	71.3%	10.1%

표 4. JHMDB(분할1), UCF101(분할1) 및 AVA에서 동작 감지를 위한 프레임 맵 @ IoU 0.5. JHMDB는 클립당 최대 40개의 프레임이 있습니다. UCF101-24의 경우, 평가를 위해 20,000개의 프레임 하위 집합을 무작위로 샘플링합니다. 우리 모델은 JHMDB와 UCF101-24에서 최첨단 성능을 얻었지만 AVA의 세분화된 특성으로 인해 어려움을 겪었습니다.

	JHMDB	UCF101-24	AVA
동작 감지	76.7%	76.3%	15.8%
배우 감지	92.8%	84.8%	75.3%

표 5. JHMDB(분할1), UCF101-24(분할1) 및 AVA 벤치마크에서 액션 감지 및 액터 디텍션 성능에 대한 프레임맵 @ IoU 0.5. 인간 어노테이터는 일관성이 있기 때문에 원자적 시각적 액션을 인식하는 데 개선할 여지가 상당히 많다는 것을 알 수 있습니다.

로캘라이제이션과 엔드투엔드 탐지 성능 간의 격차는 AVA에서 거의 60%에 달하는 반면, JHMDB와 UCF101에서는 15% 미만에 불과합니다. 이는 AVA의 주요 난이도가 로컬라이제이션이 아닌 액션 분류에 있음을 시사합니다. 그림 9는 고득점 오경보의 예시를 보여 주며, 인식의 차이

그림 9. 빨간색 상자는 흡연에 대한 고득점 오경보를 나타냅니다. 이 모델은 종종 세분화된 세부 사항을 구분하는 데 어려움을 겪습니다.

는 세분화된 세부 정보에 있음을 시사합니다.

어떤 카테고리가 어려운가요? 훈련 예제 수가 얼마나 중요할까요? 그림 8은 카테고리와 훈련 예제 수에 따른 성능을 세분화하여 보여줍니다. 일반적으로 데이터가 많을수록 성능이 향상되지만, 이상값을 보면 모든 카테고리의 복잡성이 동일하지는 않다는 것을 알 수 있습니다. 수영과 같이 장면 및 사물과 상관관계가 있는 카테고리나 넘어짐과 같이 다양성이 낮은 카테고리는 훈련 예시 수가 적음에도 불구하고 높은 성능을 얻습니다. 반대로 데이터가 많은 카테고리는 그렇지 않습니다,

만지거나 담배를 피우는 것과 같은 동작은 시각적 변화가 크거나 세밀한 식별이 필요하기 때문에 상대적으로 낮은 성능을 얻을 수 있으며, 이는 사람-사물 상호작용에 대한 연구에 동기를 부여합니다[7, 12]. 우리는 원자적인 동작을 인식하는 데 있어 이점을 얻으려면 AVA와 같은 대규모 데이터 세트뿐만 아니라 동작 및 상호 작용에 대한 풍부한 모델이 필요할 것이라고 가설을 세웠습니다.

7. 결론

이 백서에서는 15분 분량의 다양한 영화 세그먼트에서 1Hz의 원자 동작에 대한 시공간적 주석이 포함된 AVA 데이터 세트를 소개합니다. 또한 기준이 될 표준 벤치마크에서 현재 기술을 능가하는 방법을 제안합니다. 이 방법은 AVA 데이터 세트의 성능이 UCF101 또는 JHMDB에 비해 현저히 낮기 때문에 새로운 동작 인식 접근법 개발의 필요성을 강조합니다. 향후 작업에는 원자 동작을 기반으로 더 복잡한 활동을 모델링하는 것이 포함됩니다. 현재의 시각적 분류 기술로는 '식당에서 식사하기'와 같은 이벤트를 거친 장면/비디오 수준에서 분류할 수 있지만, AVA의 세밀한 시공간적 세분성에 기반한 모델은 개별 에이전트의 행동 수준에서 이해를 용이하게 합니다. 이는 인간이 무엇을 하고 있는지, 다음에 무엇을 할 수 있는지, 그리고 무엇을 할 수 있는지 이해하는 '소셜 시각 지능'을 컴퓨터에게 부여하기 위한 필수 단계입니다.

를 달성하기 위해 노력하고 있습니다.

감사의 말 이 작업에 대한 토론과 의견을 제공해 주신 Abhinav Gupta, Abhinav Shrivastava, Andrew Gallagher, Irfan Essa, Vicky Kalo-geiton에게 감사의 말씀을 전합니다.

참조

[1] S. 아부-엘-하이자, N. 코타리, J. 리, P. 낫세프, G. 토데리치, B. Varadarajan, and S. Vijayanarasimhan. YouTube- 8M: 대규모 비디오 분류 벤치마크. *arXiv:1609.08675*, 2016. 2

[2] D. 아리존. *영화 언어의 문법*. 실만-제임스 Press, 1991. 2

[3] R. 바커와 H. 라이트. *중서부와 그 아이들: 미국 마을의 심리적 생태*. 행, 피터슨 and Company, 1954. 2

[4] M. 블랭크, L. 고렐릭, E. 셰흐트만, M. 이란니, R. 바스리. 시공간 도형으로서의 동작. In *ICCV*, 2005. 2

[5] F. 카바 하일브론, V. 에스코르시아, B. 가넴, J. C. 니블스. ActivityNet: 인간의 활동성 이해를 위한 대규모 비디오 벤치마크 . In *CVPR*, 2015. 2

[6] J. 카레이라와 A. 지서만. 행동 인식의 시대는 끝났나요? 새로운 모델과 키네틱스 데이터 세트. In *CVPR*, 2017. 2, 3, 6

[7] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: 이미지에서 인간과 물체의 상호작용을 인식하기 위한 벤치마크. In *ICCV*, 2015. 3, 8

[8] K.-W. 처치와 P. 행크스. 단어 연관 규범, 뮤추얼 정보 및 어휘. *전산 언어*- 틱, 16(1), 1990. 5

[9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, 및 A. Zisserman. PASCAL Visual 객체 클래스 챌린지: 회고. *IJCV*, 2015. 3, 7

[10] 지나 데이비스 미디어 젠더 연구소. 릴 진실: 여성은 보이지도 들리지도 않는다. https://seejane.org/research-informs-empowers/data/, 2016. 3

[11] G. 그키오사리, J. 말릭. 액션 튜브 찾기. In *CVPR*, 2015. 3

[12] R. Goyal, S. E. 카호, V. 미칼스키, J. 마터진스카, S. Westphal, H. Kim, V. Haenel, I. Fru¨nd, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, 및 R. Memisevic. 시각적 상식을 학습하고 평가하기 위한 "무언가 무언가" 비디오 데이터베이스. *ICCV*, 2017. 2, 8

[13] S. 굽타 및 J. 말릭. 시각적 의미론적 역할 라벨링. *CoRR*, abs/1505.04474, 2015. 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. 이미지 인식을 위한 심층 잔여 학습 . In *CVPR*, 2016. 6

[15] G. V. 혼과 P. 페로나. 악마는 꼬리에있다: 야생에서의 세분화된 분류. 2

[16] R. Hou, C. Chen, and M. Shah. 비디오에서 동작 감지를 위한 합성곱 신경망(T-CNN). In *ICCV*, 2017. 2, 3, 6, 7

[17] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, 및 K. Murphy. 최신 컨볼 루션 물체 감지기의 속도/정확도 트레이드 오프 . In *CVPR*, 2017. 6

[18] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. 석탄카르, M. 샤. "야생에서" 동영상에 대한 동작 인식에 대한 THUMOS 챌린지. *CVIU*, 2017. 2

[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, 및 T. Brox. FlowNet 2.0: 딥 네트워크를 통한 광학 유량 추정의 진화 . In *CVPR*, 2017. 6

[20] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. Black. 행동 인식을 이해하는 병동. In *ICCV*, 2013. 2, 3, 7

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, 및 L. Fei-Fei. convo- lutional 신경망을 사용한 대규모 비디오 분류. In *CVPR*, 2014. 2

[22] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. 동역학 인간 행동 비디오 데이터 세트. 2, 6

[23] Y. Ke, R. Sukthankar, 및 M. Hebert. 체적 특징을 이용한 효율적인 시각적 이벤트 탐지. In *ICCV*, 2005. 3

[24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, 및 T. Serre. HMDB: 인간 동작 인식을 위한 대규모 비디오 데이터베이스 . In *ICCV*, 2011. 2

[25] H. W. 쿤. 헝가리식 할당 문제 해결 방법. *해군 연구 물류 (NRL)*, 2(1-2):83-97, 1955. 4

[26] M. Marszalek, I. Laptev, 및 C. Schmid. 맥락에 맞는 행동. *CVPR*, 2009. 2

[27] P. Mettes, J. van Gemert, 및 C. Snoek. Spot On: 포인트 감독 제안을 통한 액션 로컬라이제이션. *ECCV*, 2016. 3

[28] M. 몬포트, B. 저우, S. A. 바갈, T. 얀, A. 안도니안, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Von- drick 등. 순간 데이터 세트: 이벤트 이해를 위한 100만 개의 동영상. 2

[29] P. 오버, G. 아와드, M. 미셸, J. 피스커스, G. 샌더스, W. Kraaij, A. Smeaton, and G. Que´not. TRECVID 2014 - 목표, 과제, 데이터, 평가 메커니즘에 대한 개요 및 메트릭, 2014. 2

[30] X. 펭과 C. 슈미드. 동작 감지를 위한 다중 지역 2스트림 R-CNN . In *ECCV*, 2016. 3, 6, 7

[31] S. 렌, K. 허, R. 기르식, J. 선. 더 빠른 R-CNN: 지역 제안 네트워크를 통한 병동 실시간 물체 감지- 작동. In *NIPS*, 2015. 3, 4, 6

[32] M. 로드리게스, J. 아메드, 및 M. 샤. 액션 MACH: 액션 인식을 위한 시공간적 최대 평균 상관관계 높이 필터 . In *CVPR*, 2008. 2, 3

[33] S. 사하, G.싱, F. 쿠졸린. AMTnet: 엔드 투 엔드 훈련 가능한 심층 아키텍처에 의한 액션 마이크로 튜브 회귀. In ICCV, 2017. 3

[34] S. 사하, G. 싱, M. 사피엔자, P. 토르, 및 F. 쿠졸린. 동영상에서 다중 시공간 액션 튜브 감지를 위한 딥러닝 . In *BMVC*, 2016. 3

[35] C. 슐트, I. 라프테프, 및 B. 카푸토. 인간의 행동 인식: 로컬 SVM 접근법. In *ICPR*, 2004. 2

[36] G. Sigurdsson, O. Russakovsky, A. Farhadi, I. Laptev, 및 A. 굽타. 시간에 대한 많은 고민: 시간 데이터에 대한 철저한 주석. *인간 계산 컨퍼런스 및 크라우드소싱*, 2016. 4

[37] G. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, 및 A. 굽타. 가정에서 할리우드: 활동 이해를 위한 크라우드소싱 데이터 수집. In *ECCV*, 2016. 2, 6

[38] G. Singh, S. Saha, M. Sapienza, P. Torr, 및 F. Cuzzolin. 온라인 실시간 다중 시공간 행동 현지화 및 예측. In *ICCV*, 2017. 3, 7

[39] K. Soomro, A. Zamir, 및 M. Shah. UCF101: 야생의 비디오에서 101개의 인간 행동 클래스로 구성된 데이터 세트. 기술 보고서 CRCV-TR-12-01, University of Central Florida, 2012. 2, 3

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. 컴퓨터 비전을 위한 초기 아키텍처 재고. In *CVPR*, 2016. 6

[41] V.칼로게이톤, P. 와인자펠, V. 페라리, C. 슈미드. 시공간적 동작 위치 측정을 위한 작용 세뇨관 검출기. In *ICCV*, 2017. 2, 3, 6, 7

[42] L. Wang, Y. Qiao, X. Tang, and L. Van Gool. 하이브리드 완전 컨볼루션 네트워크를 이용한 행동성 추정. In *CVPR*, 2016. 7

[43] P. Weinzaepfel, Z. Harchaoui, 및 C. Schmid. 시공간적 동작 위치 추적을 위한 학습. In *ICCV*, 2015. 3

[44] P. Weinzaepfel, X. Martin, and C. Schmid. 약하게- 감독된 행동 현지화를 향하여. *arXiv:1605.05197*, 2016. 3

[45] L. 우, C. 셴, 및 A. 반 덴 헹겔. PersonNet: 심층 컨볼루션 신경망을 이용한 개인 재식별. *arXiv 사전 인쇄물 arXiv:1601.07255*, 2016. 4

[46] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, L. Fei-Fei. 매 순간이 중요합니다: 조밀하고 상세한 라벨-복잡한 비디오에서 동작의 ing. *IJCV*, 2017. 2

[47] J. Yuan, Z. Liu, and Y. Wu. 효율적인 액션 탐지를 위한 차별적 하위 볼륨 검색 . In *CVPR*, 2009. 3

[48] H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba. SLAC: 행동 분류를 위한 희소 레이블 데이터 세트 및 지역화. *arXiv preprint arXiv:1712.09374*, 2017. 2

[49] M. 졸파가리, G. 올리베이라, N. 세다갓, T. 브록스. 동작 분류 및 감지를 위해 포즈, 동작, 외모를 활용하는 연쇄 멀티스트림 네트워크. In *ICCV*, 2017. 3