AVA-키네틱스 현지화된 인간 행동 비디오 데이터 세트

앙 리[1] 메가나 토타쿠리[1] 데이비드 A. 로스[2]
Joa˜o Carreira[1] 알렉산더 보스트리코프[1*] 앤드류 지서만 [1,4]

[1]딥마인드 [2]구글 리서치 [4]VGG, 옥스퍼드

{anglili,sreemeghana,dross,joaoluis,zisserman}@google.com
alexander.vostrikov@gmail.com

초록

이 백서에서는 AVA-Kinetics의 현지화된 인간 행동 비디오 데이터 세트에 대해 설명합니다. 이 데이터 세트는 AVA 표기 프로토콜을 사용하여 Kinetics-700 데이터 세트의 비디오에 주석을 달고, 새로운 AVA 주석이 달린 Kinetics 클립으로 원본 AVA 데이터 세트를 확장하여 수집합니다. 이 데이터세트에는 키프레임의 각 인간에 대해 80개의 AVA 액션 클래스로 주석이 달린 23만 개 이상의 클립이 포함되어 있습니다. 유니티는 주석 처리 과정을 설명하고 새로운 데이터 세트에 대한 통계를 제공합니다. 또한 AVA-키네틱스 데이터 세트에서 비디오 액션 트랜스포머 네트워크를 사용한 기준선 평가를 통해 AVA 테스트 세트의 액션 분류 성능이 개선되었음을 보여줍니다. 데이터 세트는 https://research.google.com/ava/ 에서 다운로드할 수 있습니다.

1. AVA-키네틱스 데이터 세트

키네틱스[1, 6] 데이터 세트는 연구자들이 이미지넷[2] 및 이미지 표현과 유사하게 다양한 다운스트림 작업에서 미세 조정할 수 있는 아키텍처와 사전 훈련 모델을 탐색할 수 있는 대규모 분류 작업을 제공함으로써 비디오의 재현 학습을 지원하기 위해 만들어졌습니다.

다운스트림 작업은 더 상세한 주석이 필요한 경향이 있으므로 대규모로 주석을 달 때 비용이 더 많이 듭니다. AVA 데이터세트[5]는 주석 달기에 비용이 많이 드는 비디오 작업의 한 가지 영향력 있는 예로, 키네틱스에서와 같이 클립당 레이블을 지정하는 대신 프레임의 하위 집합에 있는 모든 사람에게 레이블 세트를 지정합니다. 이 작업은 키네틱스에 대한 사전 학습을 통해 작은 수준의 정확도를 얻을 수 있다는 점에서도 흥미롭습니다. 최신 기술은 여전히 평균 정확도가 34%에 불과합니다[3].

이를 계기로 두 데이터 세트의 크로스오버를 만들게 되었습니다. AVA-Kinetics 데이터 세트는 AVA 및 Kinetics-700 데이터 세트를 기반으로 하여 많은 Kinetics에 AVA 스타일의 인간 동작 및 로컬라이제이션 주석을 제공합니다.

*딥마인드에서 작업했습니다. 이제 Citadel과 함께합니다.

그림 1. AVA-Kinetics 데이터 세트의 키 프레임 예시. 이 주석에는 AVA 스타일 바운딩 박스와 그에 연관된 AVA 레이블이 포함되어 있습니다. 키 프레임은 클립 수준 레이블 '높이뛰기'로 주석이 달린 키네틱스 클립의 일부입니다.

비디오. 이 섹션에서는 AVA와 Kinetics 데이터셋을 간략하게 소개하고, AVA-Kinetics 데이터셋을 구축하기 위해 적용된 AVA 주석 절차를 설명합니다. 데이터 세트의 통계도 섹션 2에서 제공하고 분석합니다.

1.1. 배경

AVA 데이터세트[5]는 430개의 15분짜리 동영상 클립에 80개의 원자적인 시각적 동작을 조밀하게 주석으로 달았습니다. 사람의 행동은 초당 한 번 샘플링된 키 프레임에서 각 비디오의 각 사람에 대해 독립적으로 주석이 달립니다. 또한 이 데이터 세트는 각 인물을 둘러싸는 바운딩 박스를 비디오화하며, 각 인물은 총 160만 개의 레이블에 대해 함께 발생하는 모든 동작(*예:* 말하는 동안 서 있기)에 레이블이 지정되어 있습니다. 키네틱스 데이터 세트는 다양한 범위의 인간 행동을 포괄하는 또 다른 대규모 큐레이팅된 인간 행동 인식용 비디오 데이터 세트입니다. 키네틱스는 지난 몇 년 동안 키네틱스-400에서 키네틱스-700으로 발전해 왔습니다. Kinetics-700 데이터 세트에는 700개의

각 클래스당 최소 600개 이상의 클립과 총 약 65만 개의 비디오 클립으로 구성된 휴먼 액션 클래스입니다. 각 클래스의 각 클립은 서로 다른 인터넷 동영상에서 가져온 것으로, 약 10초 분량이며 동영상에서 발생하는 주요 동작을 설명하는 단일 레이블이 있습니다.

1.2. 데이터 주석 프로세스

AVA-키네틱스 데이터 세트는 AVA 스타일의 바운딩 박스와 원자 동작으로 키네틱스 데이터 세트를 확장합니다. 아래에 설명된 프레임 선택 절차를 사용하여 각 키네틱스 비디오에 대해 단일 프레임에 주석을 달 수 있습니다.

AVA 어노테이션 프로세스는 학습 데이터의 하위 집합과 Kinetics-700 데이터 세트의 검증 및 테스트 세트에 있는 모든 비디오 클립에 적용됩니다. 각 키네틱스 비디오 클립의 바운딩 박스에 주석을 다는 절차는 다음과 같습니다:

1. *사람 감지*: 사전 학습된 더 빠른 RCNN 적용 10초 길이의 비디오 클립의 각 프레임에서 [8] 사람 감지기를 사용합니다.

2. *키 프레임 선택*: 각 동영상 클립의 키 프레임으로 사람 감지 신뢰도가 가장 높은 프레임을 선택하고, 클립의 시작/종료 지점에서 최소 1초 이상 떨어진 프레임을 선택합니다.

3. *누락된 상자 주석*: 인간 어노테이터가 키 프레임에 누락된 경계 상자를 확인하고 주석을 달 수 있습니다.

4. *사람 행동 주석*: 키 프레임을 중심으로 2초 분량의 비디오 클립을 가져옵니다. 그런 다음 여러 사람(최소 3명)이 키 프레임 경계 상자에 있는 인물에 해당하는 액션 레이블을 제안합니다.

5. *사람의 작업 검증*: 인간 평가자가 제안된 모든 작업 라벨을 평가하여 최종 검증을 진행합니다. 평가자 3명 중 2명 이상의 과반수에 의해 확인된 각 라벨은 유지됩니다.

원본 AVA 데이터 세트와의 차이점은 각 키네틱스 비디오 클립에 대해 하나의 키 프레임에만 주석이 달렸다는 점입니다. 키네틱스 훈련 세트는 다음과 같이 주석을 위해 샘플링됩니다. 먼저, 기존 80개의 AVA 클래스 중 27개의 후보를 선정하여 인식 성능이 좋지 않은 AVA 클래스의 우선순위를 정합니다.[1] 27개의 후보군을 선정하여 우선순위를 정합니다. 이 후보 목록과 일치하는 텍스트를 기준으로 키네틱스 데이터 세트(부록 A에 나열됨)에서 115개의 관련 액션 클래스를 직접 선택하고, 해당 액션이 포함된 모든 키네틱스 비디오에 주석 파이프라인을 적용합니다.

[1] 우리가 우선시하는 저조한 AVA 수업은 수영하기, (물건) 밀기, (사람에게) (물건) 주기/제공하기, 옷 입기/입기, (예: TV) 보기, 던지기, 컴퓨터 작업하기, (예: 산에) 오르기, (사람에게서) (물건) 가져가기, 듣기 (예:, 음악에 맞춰), 노래 부르기(예: 자기, 사람, 그룹), (사람) 들기, (사람) 잡기, 내려놓기, 자르기, 사진 찍기, (물건) 당기기, 들어가기, 돌리기(예: 드라이버), 들기/집기, (물건) 가리키기, (다른 사람) 밀기, (물건) 치기, 쓰러지기, 촬영, 점프/뜀, 손 흔들기.

표 1. AVA-키네틱스 데이터 세트에 대한 다양한 분할의 주석이 달린 프레임 수와 고유 비디오 클립 수입니다. AVA-Kinetics 데이터는 AVA와 키네틱스의 조합입니다. 주석이 달린 프레임의 수는 AVA와 키네틱스가 거의 비슷하지만, 키네틱스가 더 많은 고유 비디오를 AVA-Kinetics 데이터 세트에 제공합니다.

	# 고유 프레임			# unique videos		
	AVA	키네틱스	AVA-키네틱스	AVA	Kinetics	AVA-Kinetics
기차	210,634	141,457	352,091	235	141,240	141,475
Val	57,371	32,511	89,882	64	32,465	32,529
테스트	117,441	65,016	182,457	131	64,771	64,902
합계	385,446	238,984	624,430	430	238,476	238,906

나머지 키네틱스 클래스의 클립은 단일 형식으로 샘플링 됩니다(아직 모든 클립에 주석을 달지 않았습니다). 트레이닝 세트와 달리 키네틱스 검증 및 테스트 세트는 모두 완전히 주석을 달았습니다.

2. 데이터 통계

이 섹션에서는 AVA-Kinetics의 데이터 분포 특성을 살펴보고 기존 AVA 데이터 세트와 비교합니다. 데이터 세트의 통계는 표 1에 나와 있으며, 이 표는 데이터 세트의 총 고유 프레임 수와 고유 비디오의 수를 보여줍니다. Kinetics 데이터 세트는 많은 수의 동영상을 포함하고 있기 때문에 AVA-Kinetics 데이터 세트에 훨씬 더 많은 고유 동영상을 제공합니다.

이 섹션에서는 80개의 클래스가 포함된 AVA v2.2의 통계를 보여드리지만, 실험에서는 AVA 챌린지의 프로토콜을 따라 60개의 클래스만 예측합니다.

2.1. 샘플 배포

각 클래스에 대한 샘플 수는 그림 2에 나와 있습니다. 하나의 샘플은 하나의 액션 레이블이 있는 하나의 바운딩 박스에 해당합니다. AVA와 키네틱스의 클래스 통계는 비슷한 경향을 보이며, 특히 두 데이터 세트 모두 롱테일 샘플 분포를 보입니다. 그러나 키네틱스가 대부분의 클래스에 상당한 수의 샘플을 추가하는 것을 관찰할 수 있습니다. 특히 '듣기' 클래스의 경우, 키네틱스가 훨씬 더 많은 훈련 샘플을 생성합니다.

2.2. 비디오 배포

그림 3은 두 데이터 세트 간의 동영상 다양성을 보여줍니다. 이 그래프는 로그 스케일로 표시되어 있으며, 각 레이블에 대해 훨씬 더 많은 고유 동영상을 포함하고 있다는 특성을 보여줍니다. 앞서 설명했듯이 Kinetics는 10초 길이의 여러 개의 고유한 짧은 동영상으로 구성되어 있으며, 하나의 키 프레임에 AVA 레이블로 주석을 달았습니다. AVA 비디오는 훨씬 더 길며 액션 로컬라이제이션을 위한 여러 클립을 제작하는 데 사용됩니다. AVA의 경우 클래스당 최대 고유 비디오 수는 235개이지만, 키네틱스는 상위 클래스의 고유 비디오 수가 10,000개가 넘습니다. 또한 키네틱스의 클래스 중 절반은 300개 이상의 고유 비디오 클립을 보유하고 있습니다.

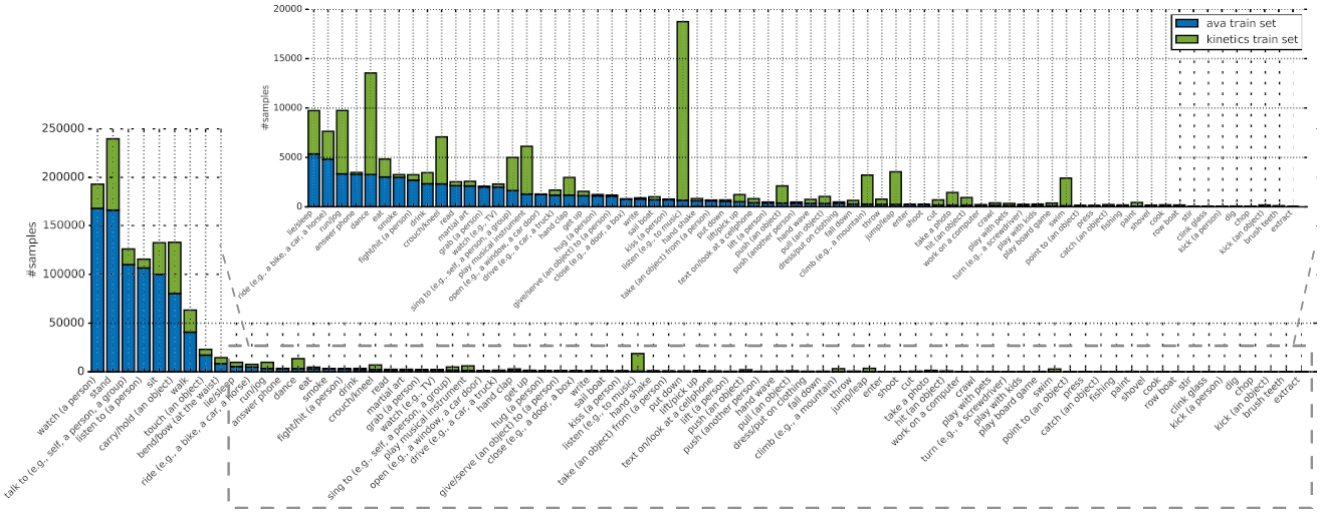

그림 2. AVA 훈련 세트(파란색)와 키네틱스 훈련 세트(녹색)를 비교한 AVA 클래스당 샘플 수입니다. 누적 막대는 AVA-키네틱스 훈련 세트의 분포를 보여줍니다. 분포는 롱테일입니다. 오른쪽 상단에 꼬리 부분이 확대되어 있습니다.

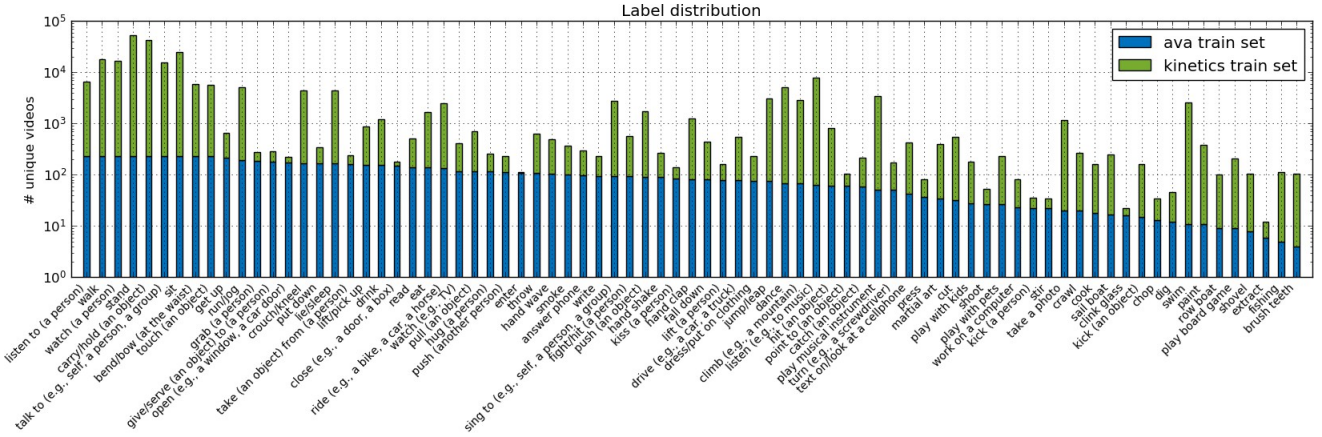

그림 3. AVA 트레인 세트(파란색)와 키네틱스 트레인 세트(녹색)를 비교한 AVA 클래스별 고유 비디오 수. 누적 막대는 AVA-키네틱스 훈련 세트의 분포를 보여줍니다. Y축은 로그 스케일입니다. 두 데이터 세트의 동영상 분포는 서로 다릅니다. AVA 데이터에는 긴 비디오가 있는 반면 Kinetics에는 짧은 비디오 클립이 있으므로 AVA의 클래스당 고유 비디오 수가 훨씬 적습니다.

2.3. 키네틱스와 AVA 클래스의 상관 관계

키네틱스 비디오에 대한 AVA 클래스와 키네틱스 클래스 모두에 대한 주석이 있으므로 두 가지 유형의 클래스에 대한 노멀라이즈드 포인트 상호 정보(NPMI)를 계산합니다. NPMI는 다음과 같이 정의됩니다.

$$NPMI(x, y) = \frac{\log p(x) + \log p(y)}{\log p(x, y)} - 1 \qquad (1)$$

여기서 $p(x)$는 클래스 x의 빈도입니다. $NPMI(x, y)$는 -1에서 1 사이의 실수 값입니다. $NPMI(-, -) = 1$은 두 클래스의 상관관계가 높음을 의미하며, $NPMI(-, -) = -1$은 두 레이블이 절대 함께 발생하지 않음을 의미합니다.

또한 각 운동학 클래스의 NPMI 점수에 따라 AVA 클래스의 순위를 매깁니다. 표 2에서 몇 가지 예시를 통해 해당 NPMI를 확인할 수 있습니다.

점수. 키네틱스 수업 "강남 스타일 춤추기"는 AVA 수업 "춤추기", "듣기(예: 음악 듣기)" 및 "보기(예: TV 보기)"와 밀접한 관련이 있습니다. '춤추다'는 직접적인 관련이 있지만, '듣다'와 '보기'는 다른 사람이 강남 스타일로 춤추는 것을 보고 있을 때 발생할 수도 있습니다.

표 3에는 역상관관계, *즉* 특정 AVA 클래스와 관련된 상위 키네틱스 클래스가 나와 있습니다. 발생 빈도가 가장 낮은 AVA 클래스 목록을 선택하고 관련성이 가장 높은 키네틱스 클래스 레이블을 표시합니다. "저어" 클래스는 "스크램블링 에그", "차 만들기", "'슬라임 만들기"와 관련이 있습니다. 세 가지 요리 활동은 모두 실제로 저어주는 동작이 필요합니다.

표 2. AVA-Kinetics 데이터 세트에서 정규화된 포인트별 상호 정보(NPMI)로 순위를 매긴 샘플 Kinetics 클래스 세트와 가장 연관성이 높은 AVA 클래스. 각 클래스 레이블 옆에는 NPMI 점수가 표시되어 있으며, 값이 높을수록 상관관계가 강하다는 것을 나타냅니다.

운동학 클래스	**관련성이 가장 높은 AVA 클래스**
사리를 입다	옷을 입다/입다(0.73); 서다(0.11); 대화하다(예: 자기, 사람, 그룹)(0.04); 휴대전화로 통
화하다(0.04).	전화 받기(0.74); 휴대폰으로 문자 보내기/보기(0.25); 운전하기(예: 자동차, 트럭
)(0.16); 양파 썰기(0.16)	자르다(0.70); 자르다(0.40); 쓰기(0.29);
양치질	양치질(0.85), (사람) 들어올리기(0.19), 휴대폰으로 문자 보내기/보기
(0.06), 항해(0.06)	범선(0.70), 낚시(0.16), 운전(예: 자동차, 트럭)(0.10);
카누 또는 카약	노 젓는 보트(0.81); 범선(0.43); 낚시(0.32);
낚싯줄 캐스팅	낚시(0.73), 회전(예: 드라이버)(0.21), 요트(0.20);
눈 삽질	삽질(0.71), 운전(예: 자동차, 트럭)(0.17), 구부리다/절다(허리)(0.15);
페인트 롤러 사용	페인트 칠하기(0.76); (물건)을 (사람에게) 주다/제공하다(0.09); (예: 자기, 사람, 그룹)과 대화하
다(0.07); 컵을 홀짝이다.	마시다(0.70); 앉다(0.22); 거짓말/수면(0.20);
물 담뱃대 흡연	연기 (0.72); 앉아 (0.14); 전화 받기 (0.11); 기어가
는 아기	기어 다니기(0.80); 거짓말/수면(0.27); 아이들과
놀기(0.19);	
춤추는 강남 스타일	춤추기(0.34), 듣기(예: 음악 듣기)(0.28), 보기(예: TV 보기)(0.21),
줄서기(0.21)	휴대폰으로 문자 보내기/보기(0.16), 전화 받기(0.15), 서 있기
(0.10);	
축구 공 패스	(물체를) 차다(0.55); 달리기/조깅(0.34); (물체를) (사람에게) 주다/서빙하다(0.14);

표 3. AVA-동역학 데이터 세트에서 정규화된 포인트별 상호 정보(NPMI)로 순위를 매긴 가장 빈도가 낮은 AVA 클래스와 관련성이 가장 높은 Kinetics 클래스. 각 클래스 레이블 옆에는 NPMI 점수가 표시되어 있으며, 값이 높을수록 상관관계가 강하다는 것을 나타냅니다.

AVA클래스	**관련성이 가장 높은 키네틱스 클래스**
닫기(예: 문, 상자)	문 닫기(0.69), 냉장고 열기(0.54), 종이 찢기(0.42), 양파 썰기(0.70),
고기 썰기(0.57), 초밥 만들기(0.41);	
잡다 (사람)	체포 (0.40); 겨드랑이 왁싱 (0.31); 다리 면도 (0.28);
수영장	청소(0.52), 쓰레기 수거하는 사람(0.45), 교회에 들어가는 모습(0.32);
들어올리기(사람)	요가(0.41); 아기 안아주기(0.36); 뒤로 젖히기(0.33);
레이저 태그	쏘기(0.43), 페인트볼 놀이(0.42), 불 피우기(0.38)키스(사람) 키
스0.69), 시계 확인(0.49), 청혼(0.37)클링크 글라스 (0.55), TV 시청(0.30), 의자에서 떨어지기	
(0.28);	
(객체)를 가리키다	일기 예보 발표(0.44), 복사(0.40), 슬롯머신(0.33), 염소젖 짜기(0.56), 굴 까기
(0.50), 젖소 젖 짜기(0.49);	

(사람)	드롭킥 (0.52); 하이킥 (0.40); 손톱 자르기 (0.38); 조개
캐기 (0.58); 땅 파기 (0.45); 나무 심기 (0.44);	
	컴퓨터 작업가죽공예(0.43); 색칠하기(0.39); 오일 교환(0.39);
	프레스풀링 에스프레소 샷(0.51); 젖소 착유(0.49); 염소 착유(0.48);
달걀	스크램블링(0.57); 차 만들기(0.54); 슬라임 만들기(0.49);

2.4. 인원 분포

그림 4는 AVA와 키네틱스 데이터 세트 모두에서 프레임당 사람 바운딩 박스의 수를 보여줍니다. 분포는 거의 동일합니다. 두 데이터 세트 모두에서 사람이 감지되지 않은 프레임이 상당수 관찰됩니다. 그리고 대부분의 키 프레임에는 단 한 명의 사람만 감지되었습니다. 5명 이상의 인물이 포함된 프레임은 매우 드뭅니다. AVA의 프레임당 평균 박스 수는 1.5개, Kinetics의 박스 수는 약 1.2개입니다.

2.5. 사람 상자 크기 분포

또한 두 데이터 세트에서 사람 바운딩 박스의 크기도 조사합니다. 그림 5는 사람 바운딩 박스 영역의 분포를 보여줍니다. 이 영역은 다음에 따라 정규화됩니다. 1x1 정사각형 이미지입니다. 대부분의 사람 상자가 이미지에 비해 작은 것을 관찰할 수 있습니다. 흥미로운 발견은 키네틱스 동영상이 AVA 동영상에 비해 사람 바운딩 박스가 더 작은 경향이 있다는 것입니다.

3. 벤치마킹 결과

우리는 지상 실측 인물 경계 상자 위에 비디오 액션 트랜스포머 네트워크[4]를 사용하여 새로운 데이터 세트를 실험합니다. 지상 실측 박스를 사용하면 객체 검출기 오류와 분리하여 동작 분류 정확도를 평가할 수 있습니다. 원래의 AVA pa- per [5]는 ResNet-50 백본과 함께 더 빠른 RCNN을 사용하여 0.5 IoU에서 평균 75%의 AP 사람 감지율을 보고하고, 액션 트랜스포머 [4]는 AVA 점수를 개선한 것으로 보고합니다.

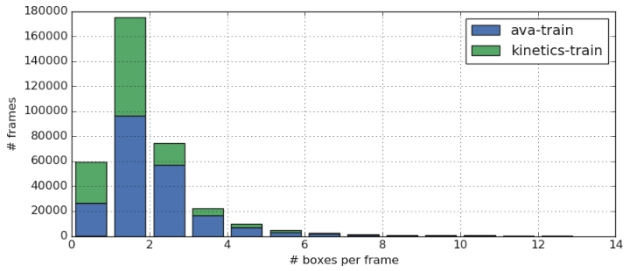

그림 4. AVA 트레인 세트(파란색)와 키네틱스 트레인 세트(녹색)를 비교한 프레임당 바운딩 박스 수입니다. 누적된 막대는 AVA-키네틱스 훈련 세트의 분포를 보여줍니다. 사람이 감지되지 않은 프레임이 상당수 있습니다. 대부분의 키 프레임에는 바운딩 박스가 하나만 포함되어 있습니다.

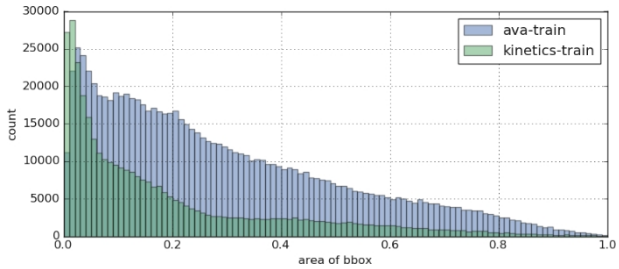

그림 5. AVA 훈련 집합(파란색)과 Kinetics 훈련 집합(녹색)에서 사람 경계 상자의 면적 분포. 면적은 1x1 정사각형에 따라 정규화됩니다. AVA의 피크 면적은 약 0.02이고 Kinetics의 피크는 약 0.01입니다. 키네틱스 데이터 세트가 훨씬 더 작은 사람 바운딩 박스를 생성하는 것을 볼 수 있습니다.

자동 감지된 바운딩 박스 대신 지상 실측 바운딩 박스를 사용할 경우 17.8%에서 29.1%로 11.2% 증가했습니다. 이는 상당한 개선이지만, 나머지 70%의 완벽한 천공률을 고려할 때 여전히 적은 수치입니다. 이러한 결과는 무엇보다도 다중 레이블의 경우 일반 동작 분류가 여전히 매우 어렵다는 것을 시사하는 것으로 보입니다(Charades [9]는 공간적 측면을 제외한 또 다른 어려운 다중 레이블 동작 분류 데이터 세트입니다). 실사 기반 행동 분류 외에도 사전 학습된 사람 감지기를 사용하여 테스트 시점에 상자를 제안하는 결과도 보고합니다.

3.1. 액션 트랜스포머: 짧은 요약

오리지널 액션 트랜스포머[4]는 3D 컨보루션 백본을 사용하여 각 클립에 대한 시공간 그리드 파운데이션을 생성합니다. 실제로 비디오는 각 키 프레임을 중심으로 여러 클립으로 나뉩니다. 시공간적 특징 그리드를 사용하여 검출기(예: Faster-RCNN의 영역 제안 네트워크)는 중간 프레임에 대한 여러 개의 오브젝트 박스를 생성합니다. 위치 임 베 딩이 특징 그리드에 추가되고, 각 박스에 대해 ROI가 풀링되며, 결과 특징이 트랜스포머 스택을 통과합니다. 트랜스포머의 쿼리

표 4. 주어진 실측 바운딩 박스와 함께 다양한 훈련/값 세트를 사용한 AVA 동작 분류에 대한 비디오 액션 트랜스포머의 성능 (mAP%).

	AVA	**동역학**	**AVA 동역학**
	val / test	val / test	val / test
AVA 열차27	.47 / 25.85	16.08 / 16.02	24.26 / 23.47
키네틱 열차	22.09 / 20.91	33.68 / 31.91	26.96 / 26.51
AVA-키네틱스 열차	**32.74 / 31.30**	**35.54 / 34.52**	**35.98 / 35.56**

표 5. 경계 박스가 제거된 다양한 훈련/값 세트를 사용한 AVA 동작 분류에 대한 비디오 액션 트랜스포머의 성능(mAP%). 모델은 실측 바운딩 박스만을 사용하여 학습됩니다.

	AVA val / test	**키네틱스** val / test	**AVA-키네틱스** val / test
AVA 열차19	.05 / 17.76	8.40 / 8.45	15.61 / 14.89
키네틱 열차	15.59 /14.05	19.05 / 18.12	16.79 / 16.35
AVA-키네틱스 열차	**23.01 / 21.23**	**20.03 / 19.74**	**22.99 / 22.70**

공간적-시간적 ROI 풀 피처를 평균화하거나 공간 구성 정보를 더 잘 보존하는 보다 정교한 병합 작업을 사용하여 얻을 수 있지만, 실측 데이터 상자 분류를 위해서는 단순한 평균화가 더 효과적이므로 이를 사용합니다. 키와 값은 ROI가 풀링되지 않은 원래의 특징 그리드에서 직접 파생됩니다. 전반적으로 모델은 각 인물과 전체 장면 사이의 관계를 포착하려고 합니다.

여기서는 재연 제안 기계가 없는 단순화된 모델을 사용합니다. 대신 지상 실측 상자에 대해 학습한 다음 지상 실측 상자 또는 최신 첨단 탐지기가 제공하는 사람 탐지에 대해 테스트합니다[10].

3.2. 지상 실측 상자를 사용한 전반적인 성능

표 4에서는 서로 다른 세 가지 훈련 세트와 세 가지 검증 세트를 사용하여 9가지 설정에서 기준값 상자를 사용한 액션 분류 성능을 보여줍니다. AVA-Kinetics 훈련/값은 기본적으로 해당 AVA와 Kinetics 데이터의 조합입니다.

표에 따르면 Kinetics 훈련 데이터를 추가하면 원래 AVA 검증 세트에서 평가할 때 성능이 향상됩니다 (+5.26mAP). 또한 Kinetics 클립에 대한 훈련이 더 안정적인 생성 기능을 제공한다는 것을 알 수 있습니다. AVA 유효성 검사에서는 AVA로 훈련된 모델에 가까운 성능을 보이지만, Kinetics 유효성 검사 클립에서는 AVA로 훈련된 모델의 성능이 훨씬 낮습니다. 마지막으로 전체 AVA-Kinetics 훈련 세트에 대해 훈련하면 예상대로 전체 AVA-Kinetics 검증 세트에서 최상의 성능을 얻을 수 있습니다.

3.3. 감지된 상자를 사용한 전체 성능

표 5는 테스트 시점에 자동 감지된 상자를 사용했을 때의 해당 결과를 보여줍니다. 중앙

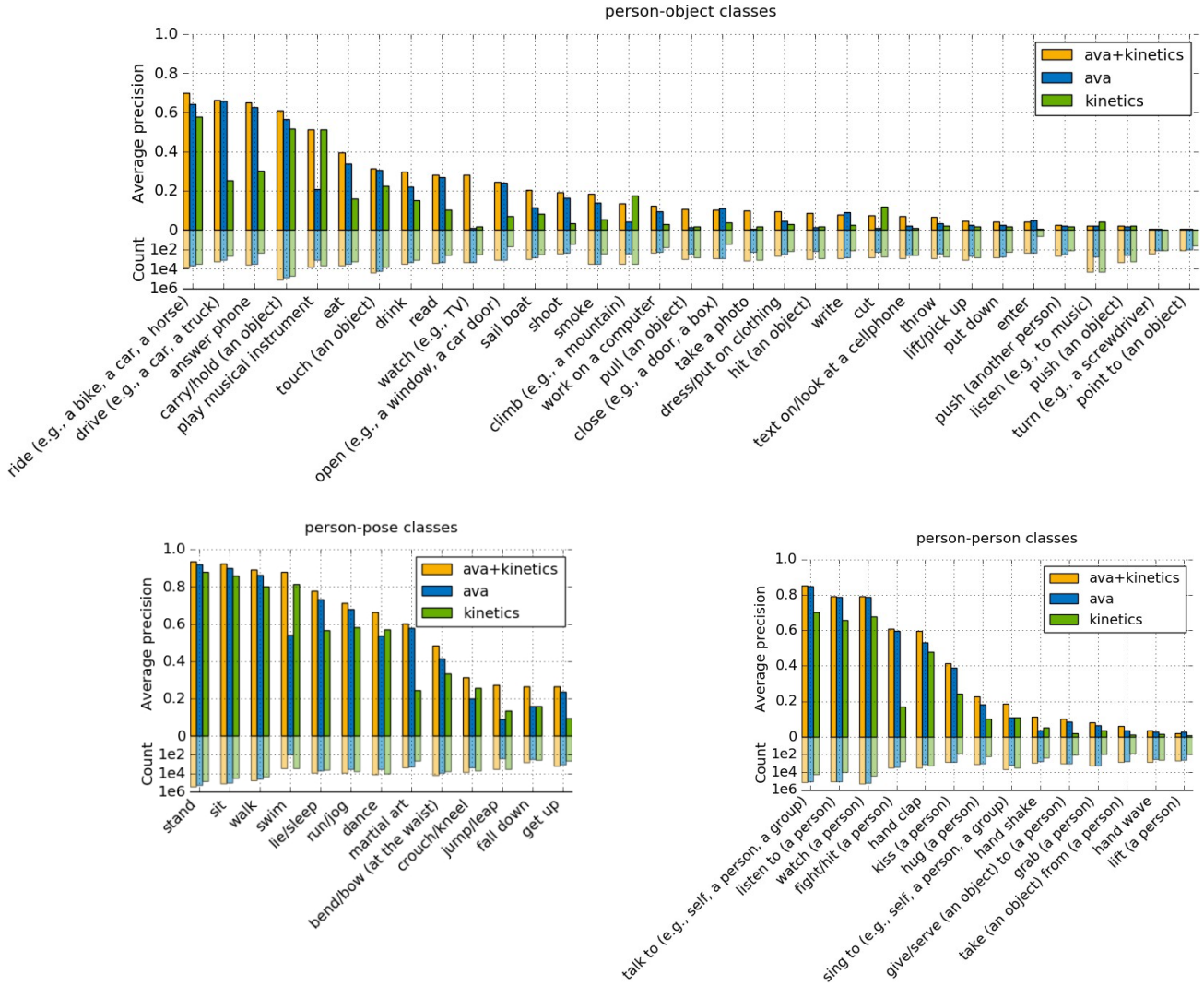

그림 6. 사람 간 상호작용, 사람-사물 간 상호작용, 사람 포즈 등 세 가지 동작 범주에 대한 AVA 검증 세트의 클래스별 성능. 0축 위의 막대는 평균 정밀도 값을 나타내고, 0축 아래의 투명한 막대는 해당 훈련 데이터 세트의 예제 수를 나타냅니다. 기준값 상자는 훈련과 테스트 모두에 사용됩니다.

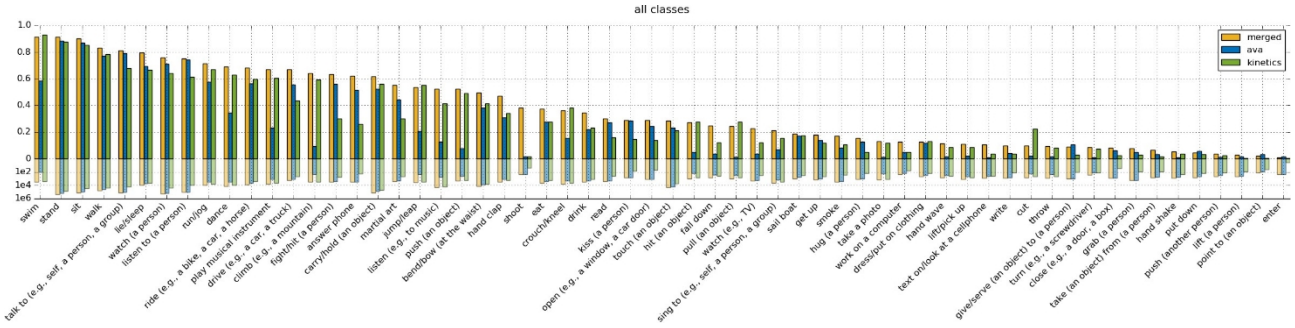

그림 7. AVA-Kinetics 검증 세트에 대한 클래스별 평가. 비교 모델은 AVA, Kinetics 및 AVA-Kinetics(결합)에 대해 훈련됩니다. 실측 데이터 상자는 훈련과 테스트 모두에 사용됩니다. 0축 위의 막대는 평균 정밀도 값을 나타내며, 0축 아래의 투명한 막대는 해당 훈련 데이터 세트의 예제 수를 나타냅니다.

Net [10] 사람 감지기를 사용하여 비디오 액션 트랜스포머 모델에 공급되는 신뢰도 높은 감지 세트(키 프레임당 약 2.5개의 박스)를 생성합니다. 객체 감지기는 COCO [7] 2017 훈련 세트에 대해 훈련되며 COCO 2017 검증 세트에서 43%의 mAP를 달성합니다. "사람" 클래스로 감지된 상자만 비디오 액션 감지를 위한 제안으로 사용됩니다. 단순화를 위해 지상 실측 상자에 대해 학습된 동일한 모델을 사용했으며, 감지된 상자가 사용된 테스트 시점에만 변경되었습니다.

결과는 대부분 실측 상자에 대한 평가 결과와 일치하며, AVA-Kinetics 훈련 세트(+3.96mAP)로 훈련할 때 AVA 검증이 개선된 것으로 나타났습니다. 불완전한 감지로 인해 전반적으로 값이 약간 낮습니다. 원래의 액션 트랜스포머 모델은 AVA-값 mAP에 대해 더 높은 AVA-트레이닝을 보고하지만 영역 제안 네트워크와 배경 네거티브 예제 선택이 포함된 더 복잡한 훈련 절차를 사용하는 반면, 우리는 훈련에 지상 진실 상자만 사용합니다.

3.4. 클래스별 성능

전체 비교 외에도 세 가지 다른 훈련 분할(AVA, 운동학, AVA- 운동학)을 사용하여 동일한 AVA 검증 세트에 대한 클래스별 성능을 플롯합니다. 그림 6에서 모든 AVA 클래스를 사람-사물 상호작용, 사람-포즈, 사람-사람 상호작용의 세 가지 범주로 나누어 비교한 결과를 확인할 수 있습니다. 총 훈련 샘플 수도 각 하위 그림에서 0축 아래의 로그 스케일로 플롯되어 있습니다. 사람-포즈 클래스에 대한 전반적인 성능이 상대적으로 높은 반면, 사람-물체 클래스가 가장 인식하기 어려운 경우인 것으로 보입니다. 특히 '보기(예: TV 시청)', '자르기', '악수', '점프/도약', '수영'에 대한 성능은 새로운 키네틱스 데이터를 사용함으로써 크게 향상되었습니다. 전체 AVA-Kinetics 검증 세트에 대한 유사한 클래스별 평가는 그림 7에 나와 있습니다.

3.5. 성능 개선 대 데이터 증가

그림 8에서는 성능 개선이 데이터 증가와 어떤 관련이 있는지 살펴봅니다. *X축은* AVA 훈련 세트에서 AVA-키네틱스 훈련 세트로 샘플 크기가 증가한 비율을 나타냅니다. 이는 기본적으로 동일한 라벨 클래스에서 키네틱스 샘플 크기와 AVA 샘플 크기의 비율입니다. Y축은 mAP의 개선도, *즉* AVA-Kinetics 세트에서 훈련된 모델의 mAP에서 AVA 세트에서 훈련된 모델의 mAP를 뺀 값입니다. 일부 클래스 이름은 해당 데이터 포인트 근처에 표시됩니다. mAP 메트릭은 AVA-Kinetics 검증 세트에서 측정됩니다. "입력"(빨간색) 클래스 하나만 성능 저하(0.76% mAP 감소)가 있는 것으로 관찰됩니다. 다른 모든 클래스는 더 많은 키네틱스 비디오에 대한 훈련을 통해 개선되었습니다. 가장 개선된 클래스 중에는 "악기 연주", "수영", "'듣기(예: 음악 듣기)"가 있습니다. - 이 경우 훈련 예제 수가 기존보다 두 배 이상 증가했습니다.

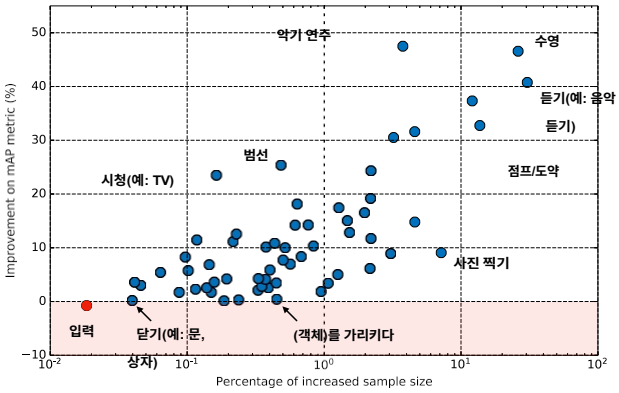

그림 8. AVA-키네틱스 검증 세트에서 mAP 성능 개선. X축은 AVA 훈련 세트의 샘플 크기 대비 추가된 키네틱스 샘플 크기의 백분율을 나타냅니다. 각 점은 하나의 클래스에 해당합니다. 파란색 점은 mAP가 개선된 클래스를 나타내고, 빨간색 점 하나만 mAP가 감소한 '입력' 클래스를 나타냅니다. 그러나 mAP 감소(-0.76%)는 0에 약간 못 미치는 미미한 수준입니다. 테스트 시점의 실측 상자를 사용하여 얻은 결과입니다.

4. 결론

유니티는 키네틱스와 AVA 데이터셋을 결합한 AVA-키네틱스 로컬라이즈드 휴먼 액션 비디오 데이터셋을 선보였습니다. AVA는 모든 사람에 대해 상세한 다중 레이블 주석을 달았지만 약 500개의 고유 동영상으로 시각적 다양성이 제한적이었던 반면, 키네틱스는 동영상당 단일 레이블을 사용하지만 60만 개의 고유 동영상으로 시각적 다양성이 매우 넓습니다. 각 키네틱스 비디오의 한 프레임에 AVA 박스와 레이블을 주석으로 달면 더 풍부한 훈련 세트와 실제 모델 일반화를 더 잘 반영하는 테스트 세트가 포함된 AVA-Kinetics 데이터 세트를 얻을 수 있습니다.

이 데이터 세트는 비디오의 다양한 환경에서 연구에 유용하게 활용될 수 있습니다. 예를 들어, 멀티태스크 학습(AVA와 키네틱스 레이블 모두에 대한 학습과 개별 레이블에 대한 학습) 또는 전이 학습(키네틱스에 대한 사전 학습 후 AVA에 대한 미세 조정은 AVA-키네틱스에 대한 직접 학습과 어떻게 비교되는가?)에 활용될 수 있습니다.

감사의 말씀:

이 데이터 세트의 수집은 딥마인드와 구글 리서치의 지원을 받았습니다. 저자들은 객체 감지에 도움을 준 비그네쉬 비로드카르, 유후이 첸, 조나단 황, 비벡 라토드, 액션 주석 파이프라인에 도움을 준 예칭 리, 논문 초안을 검토해 준 장 밥티스트 알레이락에게 감사의 말을 전합니다. 또한 이 작업에 기여해 주신 Rahul Sukthankar, Victor Gomes, Ellen Clancy, Chloe Rosenberg에게도 감사의 말씀을 전합니다.

참조

[1] 조아오 카레이라, 에릭 놀란드, 클로이 힐리어, 앤드류 지스만. 키네틱스-700 인간 행동에 대한 짧은 메모

데이터 세트. *arXiv preprint arXiv:1907.06987*, 2019.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, 및 L. Fei-Fei. ImageNet: 대규모 계층적 이미지 데이터베이스. In *CVPR09*, 2009.

[3] 크리스토프 페이첸호퍼, 하오치 판, 지텐드라 말릭, 카이밍 허. 비디오 인식을 위한 슬로우패스트 네트워크. *IEEE 국제 컴퓨터 비전 컨퍼런스 논문집, puter Vision*, 6202-6211페이지, 2019.

[4] 로힛 거드하르, 조아오 카레이라, 칼 도어쉬, 앤드류 지스-서먼. 비디오 액션 트랜스포머 네트워크. *IEEE 컴퓨터 비전 및 패턴 컨퍼런스* 논문집 *인식*, 244-253페이지, 2019.

[5] 구춘휘, 첸 선, 데이비드 로스, 칼 본드릭, 캐롤라인 판토파루, 예칭 리, 수드헨드라 비자야-나라심한, 조지 토데리치, 수산나 리코, 라훌 석-감사, 코델리아 슈미드, 지텐드라 말릭. Ava: 시공간적으로 국지화된 원자 시각적 동작의 비디오 데이터 세트. *IEEE 컴퓨터 비전 컨퍼런스 논문집 및 패턴 인식*, 6047-6056페이지, 2018.

[6] 윌 케이, 조아오 카레이라, 카렌 시모니안, 브라이언 장, 클로이 힐리어, 수드헨드라 비자야나라심한, 파비오 비올라, 팀 그린, 트레버 백, 폴 나세프, 무스타파 술레이만, 앤드류 지서먼. 동역학 인간 행동 비디오 데이터 세트. *arXiv 사전 인쇄물 arXiv:1705.06950*, 2017.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dolla´r, C. Lawrence Zitnick. 마이크로소프트 코코: 컨텍스트에서 공통 객체. *유럽 컴퓨터 비전 컨퍼런스*, 740-755 페이지. Springer, 2014.

[8] 샤오칭 렌, 카이밍 허, 로스 거식, 지안 선. 더 빠른 r-cnn: 지역 제안 네트워크를 통한 실시간 객체 감지를 향하여, 2015.

[9] 군나르 시구르드손, 구엘 바롤, 샤오롱 왕, 알리 파르하디, 이반 랩테프, 아브히나브 굽타. 집 안의 할리우드: 활동 부족을 위한 크라우드소싱 데이터 수집. *유럽 컴퓨터 비전 컨퍼런스*, 510-526페이지. Springer, 2016.

[10] 싱이 저우, 데쿠안 왕, 필립 크라웬부엘. 점으로서의 오브젝트. *arXiv 사전 인쇄물 arXiv:1904.07850*, 2019.

A. 완전 주석이 달린 동역학 수업

위에서 언급했듯이, 저희는 가장 어려운 AVA 클래스와 가장 연관성이 높은 115개의 키네틱스 클래스를 직접 골라 모든 트레이닝 세트 예제에 주석을 달았습니다. 전체 목록은 다음과 같습니다:

수영, 수영 배영, 수영 평영, 수영 나비 스트로크, 수영 앞쪽 크롤링, 돌고래와 수영, 상어와 수영, 스쿠버 다이빙, 다이빙 절벽, 헬멧 다이빙, 자동차 밀기, 카트 밀기, 수레 밀기, 휠체어 밀기, 상 주고 받기, 사리 입기, 신발 신기, TV 시청, 카드 던지기, 야구공 잡거 나 던 지 기 , 프리즈비 잡거나 던지기, 소프트볼 잡거나 던지기, 해머 던지기,

창 던지기, 도끼 던지기, 공 던지기(야구나 미식축구 제외), 투호 던지기, 칼 던지기, 눈덩이 던지기, 투정 부리기, 물풍선 던지기, 조립 블링 컴퓨터, 밧줄 오르기, 사다리 오르기, 나무 오르기, 빙벽 등반, 등산(운동), 암벽 등반, 상 주고 받기, 청혼, 헤드뱅잉, 헤드폰으로 듣기, 사일런트 디스코, 비트박스, 교회에서 복음 사인하기, 노래 부르기, 노래방, 아기 안기, 턱걸이, 팔굽혀펴기, 레슬링, 체포, 팔씨름, 레슬링, 악수, 탱고 춤, 청소 및 저크, 카드 거래, 레고 만들기, 카드 쌓기, 컵 마시기, 수박 자르기, 파인애플 자르기, 오렌지 자르기, 케이크 자르기, 사과 자르기, 사진 촬영, 줄 당기기(게임), 교회 들어가기, 렌치 사용하기, 연필 깎기, 페인트 롤러 사용하기, 데드리프트, 모자 들기, 역기 들기, 주사위 쌓기, 수화 통역, 휠체어 밀기, 압정, 미식축구, 모쉬 피트 댄스, 레슬링, 사람 때리기(복싱), 야구 타격, 샌드백, 골프 치기, 골프 드라이브, 골프 퍼팅, 부싯돌 두드리기, 자전거에서 떨어지기, 의자에서 떨어뜨리기, 페이스 플랜팅, 세발뛰기, 멀리뛰기, 높이뛰기, 번지점프, 파쿠르, 체조 텀블링, 레이저 태그 놀이, 페인트 공 놀이, 세발뛰기, 높이뛰기, 멀리뛰기, 잭 점프, 번지점프, 스프링보드 다이빙, 트램폴린에서 튕기기, 체조 텀블링, 소파 점프, 수영장 뛰어들기, 손 흔들기, 손가락 스냅, 손가락 드럼 연주, 손박수 게임, 주먹 펌핑하기