딥 비디오 액션 인식을 위한 백본 서베이

요약

중산대학교와 북경대학교의 연구원들은 비디오 행동 인식을 위한 딥 신경망 백본에 대한 포괄적인 조사를 제공하며, 아키텍처를 Two-Stream 네트워크, 3D 컨볼루션 네트워크 및 트랜스포머 기반 방법으로 분류했습니다. 이 조사는 이들의 진화와 성능 경향을 자세히 설명하며, 트랜스포머 기반 모델이 계산 복잡성 증가에도 불구하고 일반적으로 더 높은 정확도를 달성한다는 점을 지적합니다.

Table Of Contents

1. 서론

비디오 행동 인식은 컴퓨터 비전 분야의 근본적인 문제로 떠올랐으며, 감시, 로봇 공학, 인간-컴퓨터 상호작용, 그리고 점차적으로는 인터랙티브 메타버스에까지 응용되고 있습니다. 이미지 분류와 달리 비디오 행동 인식은 공간적 특징(어떤 객체가 존재하는지)과 시간적 동역학(시간이 지남에 따라 어떻게 움직이는지)을 모두 이해해야 합니다. 이러한 복잡성으로 인해 시공간 정보를 포착하도록 특별히 설계된 다양한 딥러닝 아키텍처가 개발되었습니다.

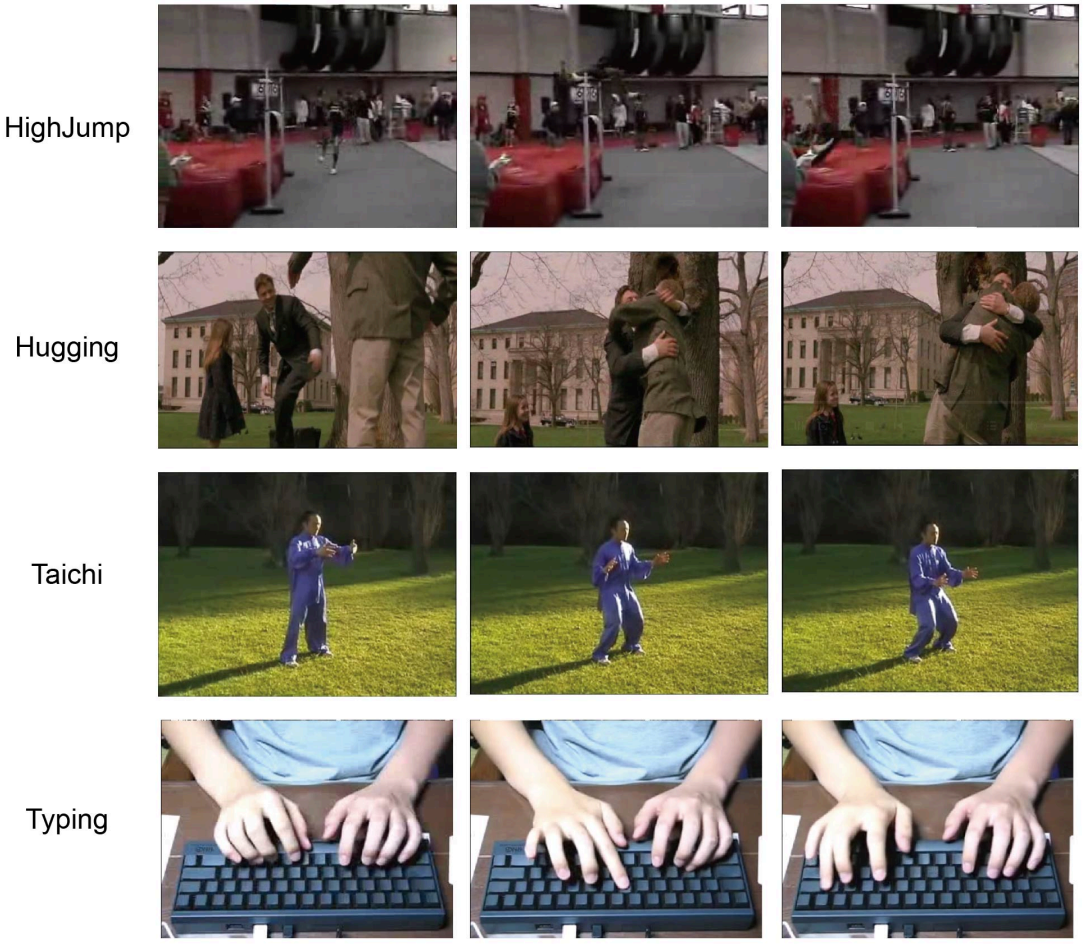

그림 1: 높이뛰기, 포옹하기, 태극권, 타이핑 등 비디오 데이터셋의 다양한 행동 예시. 각 행은 여러 프레임에 걸친 행동의 시간적 진행을 보여줍니다.

중산대학교와 베이징대학교 연구자들이 작성한 이 조사는 비디오 행동 인식을 위한 딥 신경망 백본에 대한 포괄적인 개요를 제공합니다. 이 논문은 이러한 아키텍처를 세 가지 주요 그룹으로 분류합니다: 투 스트림 네트워크(Two-Stream Networks), 3D 컨볼루션 네트워크(3D CNNs), 그리고 트랜스포머 기반 방법(Transformer-based Methods). 이 조사는 연구자와 실무자들이 비디오 이해 기술의 점점 더 복잡해지는 환경을 탐색하고 미래 연구를 위한 유망한 방향을 식별하는 데 도움을 주는 것을 목표로 합니다.

2. 비디오 행동 인식의 진화

비디오 행동 인식 분야는 지난 10년 동안 상당한 발전을 거듭했습니다. 초기 접근 방식은 HOG(Histograms of Oriented Gradients) 및 MBH(Motion Boundary Histograms)와 같은 수작업 특징에 의존했습니다. 이러한 방법은 당시에는 혁신적이었지만, 복잡한 시공간 관계를 포착하는 능력에는 한계가 있었습니다.

딥러닝 혁명은 비디오 행동 인식에 패러다임 전환을 가져왔습니다. 처음에는 이미지 인식을 위해 설계된 컨볼루션 신경망(CNN)이 비디오 데이터를 처리하도록 적용되었습니다. 이는 모션 및 시간적 의존성을 효과적으로 모델링할 수 있는 특수 아키텍처의 개발로 이어졌습니다:

1. **투 스트림 네트워크(Two-Stream Networks)** (2014): 공간 정보와 시간 정보를 분리된 스트림으로 사용하는 방식을 개척했습니다.
2. **3D CNN(3D Convolutional Networks)** (2015-2019): 2D 컨볼루션을 시간 차원으로 확장했습니다.
3. **트랜스포머 기반 방법(Transformer-based Methods)** (2020-현재): 비디오 이해에 어텐션(attention) 메커니즘을 적용했습니다.

이러한 진화는 인간의 행동을 정의하는 복잡한 시공간 패턴을 모델링하는 보다 효과적인 방법을 지속적으로 추구하는 것을 반영합니다.

3. 인기 벤치마크

비디오 행동 인식 기술의 발전은 점점 더 도전적인 벤치마크 데이터셋에 의해 주도되었습니다. 이 조사는 특히 영향력 있는 세 가지 벤치마크를 강조합니다:

- **HMDB-51**: 51개 행동 카테고리로 나뉘어 6,849개의 클립을 포함합니다. 현대 기준으로 상대적으로 작지만, 중요한 테스트베드 역할을 해왔습니다.
- **Kinetics-400**: 400개의 인간 행동 클래스를 포함하는 약 24만 개의 비디오 클립으로 구성된 대규모 데이터셋입니다. 이는 행동 인식 모델 평가를 위한 사실상의 표준이 되었습니다.
- **Something-Something V2**: 174개 행동 클래스에 걸쳐 220,847개의 비디오를 포함합니다. 시간 순서에 관계없이 행동을 인식하는 데 중점을 둔 Kinetics와 달리, Something-Something은 시간적 추론과 객체가 상호작용하는 방식 이해를 강조합니다.

이러한 벤치마크는 공간적 추론과 시간적 추론에 대한 강조점이 다르며, 모델 성능 평가를 위한 상호 보완적인 지표를 제공합니다.

4. 투 스트림 네트워크

투 스트림 네트워크는 비디오 행동 인식에 대한 초기 성공적인 딥러닝 접근 방식 중 하나를 나타냅니다. 이름에서 알 수 있듯이, 이 네트워크는 두 가지 다른 유형의 입력을 처리합니다:

1. **공간 스트림(Spatial Stream)**: 개별 RGB 프레임을 처리하여 외관 정보를 포착합니다.

2. **시간 스트림(Temporal Stream)**: 광학 흐름 필드를 처리하여 움직임 정보를 포착합니다.

Simonyan과 Zisserman이 제안한 기초적인 투 스트림 컨볼루션 네트워크는 이러한 상호 보완적인 정보 소스를 결합하는 것이 인식 정확도를 크게 향상시킨다는 것을 입증했습니다.

원래 아키텍처를 개선하기 위해 몇 가지 변형이 제안되었습니다.

- **시간 세그먼트 네트워크 (TSN)**: 비디오를 세그먼트로 나누고 희소 샘플링을 수행하여 장거리 시간 구조 모델링 문제를 해결합니다.
- **시간 관계 네트워크 (TRN)**: 여러 시간 스케일에서 프레임 간의 시간적 관계를 명시적으로 모델링합니다.
- **ActionVLAD**: 지역적으로 집계된 특징 벡터(Vector of Locally Aggregated Descriptors) 접근 방식을 사용하여 프레임 전체에 걸쳐 특징 집계를 향상시킵니다.

기본 투 스트림 네트워크의 수학적 공식은 다음과 같이 표현될 수 있습니다.

$$S_{final} = \alpha S_{spatial}(I) + (1 - \alpha) S_{temporal}(O)$$

여기서 $S_{spatial}(I)$와 $S_{temporal}(O)$는 각각 입력 프레임 I와 광학 흐름 O에 대한 공간 스트림과 시간 스트림의 점수이며, α는 융합 가중치입니다.

5. 3D 컨볼루션 네트워크

3D 컨볼루션 네트워크는 컨볼루션 연산에 시간 차원을 통합하여 기존 2D CNN을 확장합니다. 3D 컨볼루션 커널은 $K \in \mathbb{R}^{t \times d \times d}$로 표현될 수 있으며, 여기서 t는 시간적 범위이고 d는 공간적 차원입니다.

3D 컨볼루션 연산은 다음과 같이 정의됩니다.

$$V_{xyz}^{ij} = \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} w_{pqr}^{ij} \cdot v_{(x+p)(y+q)(z+r)}^{(i-1)}$$

여기서 V_{xyz}^{ij}는 i번째 레이어의 j번째 특징 맵에서 위치 (x, y, z)의 값을 나타내고, w_{pqr}^{ij}는 커널의 (p, q, r) 위치에서의 가중치입니다.

이 범주의 주요 아키텍처는 다음과 같습니다.

- **C3D**: 최초의 3D CNN 아키텍처 중 하나로, 네트워크 전체에 걸쳐 3×3×3 커널을 사용하는 3D 컨볼루션을 특징으로 합니다.
- **확장된 3D 컨브넷 (I3D)**: 2D CNN 필터를 3D로 확장하여 성공적인 이미지 분류 아키텍처 및 사전 학습된 가중치 재사용을 가능하게 합니다.
- **ResNet3D 및 ResNeXt3D**: 잔차 학습 프레임워크를 3D 컨볼루션으로 확장하여 더 깊은 네트워크를 허용합니다.

- **SlowFast 네트워크**: 서로 다른 프레임 속도로 작동하는 두 가지 경로를 사용하여 미세한 시간적 세부 정보(빠른 경로)와 의미 정보(느린 경로)를 캡처합니다.

3D CNN은 투 스트림 네트워크에 비해 시공간 모델링에 더 통합된 접근 방식을 제공하지만, 계산 복잡성이 증가하는 단점이 있습니다.

6. 트랜스포머 기반 방법

트랜스포머 기반 아키텍처는 최근 비디오 동작 인식을 위한 CNN 기반 접근 방식에 대한 강력한 대안으로 부상했습니다. 원래 자연어 처리를 위해 개발된 트랜스포머는 시퀀스 요소 간의 관계를 모델링하기 위해 어텐션 메커니즘에 의존합니다.

트랜스포머의 셀프 어텐션 메커니즘은 다음과 같이 공식화될 수 있습니다.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

여기서 Q, K, V는 각각 쿼리, 키, 값 행렬이며, d_k는 키의 차원입니다.

비디오 동작 인식을 위한 주목할 만한 트랜스포머 기반 접근 방식은 다음과 같습니다.

- **비디오 트랜스포머 네트워크 (VTN)**: CNN 백본에서 추출된 비디오 토큰에 트랜스포머를 적용합니다.
- **TimeSformer**: 시공간 관계를 효율적으로 모델링하기 위해 분할된 시공간 어텐션을 사용합니다.
- **비디오 스윈 트랜스포머**: 3D 이동 윈도우를 통합하여 계층적 스윈 트랜스포머 아키텍처를 비디오 데이터로 확장합니다.
- **MViT (멀티스케일 비전 트랜스포머)**: 공간 해상도를 점진적으로 줄이면서 채널 차원을 늘리는 멀티스케일 아키텍처를 사용합니다.
- **Side4Video**: 사전 학습된 이미지 모델을 비디오 인식에 적용하기 위해 사이드 적응 네트워크를 도입하여 성능과 파라미터 효율성 사이의 균형을 잘 맞춥니다.

트랜스포머 기반 방법은 장거리 의존성 모델링에 뛰어나며, 이는 장기간에 걸쳐 전개되는 동작을 이해하는 데 특히 유용합니다.

7. 성능 비교

본 설문조사는 표준 벤치마크에서의 성능을 기반으로 다양한 백본 아키텍처를 포괄적으로 비교합니다. 주요 관찰 내용은 다음과 같습니다:

1. **정확도 추세**: 트랜스포머 기반 방법은 일반적으로 CNN 기반 접근 방식보다 더 높은 정확도를 달성하며, VideoSwin 및 MViTv2와 같은 모델은 Kinetics-400에서 새로운 최첨단 결과를 기록했습니다.

2. **매개변수 효율성**: 트랜스포머는 정확도 측면에서 CNN을 종종 능가하지만, 일반적으로 더 많은 매개변수를

필요로 합니다. Side4Video는 특히 매개변수 효율적인 접근 방식으로 두드러집니다.

3. **데이터셋 차이**: 성능은 데이터셋에 따라 다르며, 일부 모델은 Kinetics(외형 인식에 중점을 둠)에서는 좋은 성능을 보이지만 Something-Something(시간적 추론이 필요함)에서는 어려움을 겪습니다.

4. **트레이드오프**: 모델 복잡성과 성능 사이에는 명확한 트레이드오프가 있으며, 더 큰 모델은 일반적으로 더 높은 정확도를 달성하지만 더 많은 컴퓨팅 자원을 필요로 합니다.

이러한 비교는 비디오 동작 인식에 있어 만능 해결책은 없으며, 백본의 최적 선택은 애플리케이션의 특정 요구사항과 데이터셋의 특성에 따라 달라진다는 점을 강조합니다.

8. 과제 및 미래 방향

상당한 진전에도 불구하고 비디오 동작 인식 분야에는 몇 가지 과제가 남아 있습니다:

1. **연산 효율성**: 비디오 처리는 본질적으로 컴퓨팅 집약적이므로 특히 자원 제약이 있는 장치에서 실시간 동작 인식이 어렵습니다.

2. **시간 모델링**: 장범위 시간 의존성을 효과적으로 포착하는 것은 여전히 어렵습니다. 특히 장기간에 걸쳐 전개되는 복잡한 동작의 경우 더욱 그렇습니다.

3. **데이터 요구사항**: 딥러닝 모델은 일반적으로 많은 양의 레이블링된 데이터를 필요로 하는데, 이는 비디오 애플리케이션의 경우 수집 비용이 많이 들고 시간이 많이 소요될 수 있습니다.

4. **도메인 적응**: 한 데이터셋에서 훈련된 모델은 다른 도메인이나 환경에 적용될 때 종종 성능이 저하됩니다.

향후 연구를 위한 유망한 방향은 다음과 같습니다:

1. **하이브리드 아키텍처**: CNN과 트랜스포머의 강점을 결합하여 더 적은 매개변수로 더 나은 성능을 달성합니다.

2. **자기 지도 학습**: 레이블이 지정되지 않은 비디오 데이터를 활용하여 명시적인 감독 없이 유용한 표현을 학습합니다.

3. **효율적인 어텐션 메커니즘**: 비디오 데이터 처리를 위한 표준 어텐션에 비해 더 효율적인 대안을 개발합니다.

4. **다른 양상과의 통합**: 시각 정보와 오디오, 텍스트 또는 기타 센서 데이터를 결합하여 더욱 견고한 동작 인식을 구현합니다.

5. **인터랙티브 메타버스에 적용**: 동작 인식이 가상 환경에서 더 자연스럽고 반응적인 상호 작용을 어떻게 가능하게 할 수 있는지 탐구합니다.

9. 결론

이 설문조사는 비디오 동작 인식을 위한 딥러닝 백본의 진화와 현재 상태에 대한 포괄적인 개요를 제공합니다. 초기 Two-Stream 네트워크부터 현재 트랜스포머 기반 아키텍처의 지배에 이르기까지, 이 분야는 인간의 동작을 정의하는 복잡한 시공간 패턴을 효과적으로 포착할 수 있는 모델을 개발하는 데 놀라운 발전을 이루었습니다.

비교 분석은 트랜스포머 기반 방법이 현재 최첨단 성능을 달성하고 있지만, 컴퓨팅 요구사항 및 매개변수 효율성 측면에서 고려해야 할 중요한 트레이드오프가 있음을 강조합니다. 백본 아키텍처의 선택은 애플리케이션의 특정 요구사항과 데이터셋의 특성에 따라 결정되어야 합니다.

이 분야가 계속 진화함에 따라, 컴퓨팅 효율성, 시간 모델링 및 데이터 요구사항과 관련된 과제를 해결하는 것은 비디오 동작 인식 기술의 광범위한 채택을 가능하게 하는 데 중요할 것입니다. 이러한 기술을 인터랙티브 메타버스와 같은 신흥 애플리케이션에 통합하는 것은 미래 연구 및 개발을 위한 흥미로운 기회를 제공합니다.

현재 접근 방식에 대한 구조화되고 포괄적인 개요를 제공함으로써, 이 조사는 비디오 이해 분야의 연구원 및 실무자에게 귀중한 자료가 되며, 액션 인식에 더욱 효과적이고 효율적인 모델을 개발하기 위한 미래 노력을 안내하는 데 도움을 줍니다.

10. 관련 인용

Karen Simonyan 및 Andrew Zisserman, “ **Two-Stream Convolutional Networks for Action Reco…** ,” Advances in neural information processing systems, vol. 27, 2014.

- 이 인용문은 동작 인식을 위한 기반이 되는 투스트림 컨볼루션 신경망 아키텍처를 소개하며, 이는 본 논문에서 논의된 많은 방법의 핵심 개념이자 기준점으로 사용됩니다.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, 및 Manohar Paluri, “ **Learning Spatiotemporal Features with 3D Convoluti…** ,” inProceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

- 이 인용문은 동작 인식을 위한 3D 컨볼루션 신경망(CNN)을 소개하며, 이는 본 논문에서 철저히 검토되고 비교된 핵심 아키텍처 군입니다.

Joao Carreira 및 Andrew Zisserman, “ **Quo Vadis, Action Recognition? A New Model and the …** ,” inproceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

- 이 연구는 본 논문에서 검토된 3D CNN 기반 동작 인식의 개발 및 평가에 필수적인 Inflated 3D(I3D) 모델과 Kinetics 데이터셋을 소개합니다.

Gedas Bertasius, Heng Wang, 및 Lorenzo Torresani, “비디오 이해에 시공간 어텐션만 있으면 되는가,”arXiv preprint arXiv:2102.05095, vol. 2, no. 3, pp. 4, 2021.

- 이 인용문은 시공간 어텐션을 사용하여 비디오 이해를 위한 TimeSformer 모델을 제안하며, 이는 논의된 세 번째 주요 아키텍처 그룹인 트랜스포머 기반 방법의 핵심 측면입니다.