행동 인식, 어디로 향하는가? 새로운 모델 및 Kinetics 데이터셋

요약

DeepMind와 옥스퍼드 대학의 연구원들은 인간 동작 인식을 위한 대규모 비디오 벤치마크인 Kinetics 데이터셋을 소개하고 Inflated 3D ConvNet(I3D) 아키텍처를 개발했습니다. 그들의 연구는 Kinetics에서 사전 훈련된 I3D가 UCF-101 및 HMDB-51과 같은 표준 동작 인식 벤치마크에서 새로운 성능 수준을 확립했음을 보여주었습니다.

Table Of Contents

1. 서론

비디오 이해는 오랫동안 컴퓨터 비전 분야에서 가장 어려운 문제 중 하나였으며, 인간의 행동을 인식하기 위해 모델이 공간 및 시간 정보를 동시에 처리해야 합니다. 이미지 분류는 ImageNet과 같은 대규모 데이터셋 및 심층 컨볼루션 신경망을 통해 놀라운 성공을 거두었지만, 비디오 행동 인식은 상대적으로 작은 데이터셋과 제한된 아키텍처 혁신에 의해 제약되어 왔습니다. 본 논문은 대규모 비디오 데이터셋과 이미지 분류 모델의 성공을 비디오 이해에 활용하는 새로운 딥러닝 아키텍처를 모두 도입함으로써 이러한 근본적인 한계를 해결합니다.

그림 1: Kinetics 데이터셋의 예시 프레임으로, 행동 인식을 위해 사용된 실제 비디오 콘텐츠의 다양하고 도전적인 특성을 보여줍니다.

저자들은 두 가지 중요한 질문을 다룹니다. 첫째, 비디오 이해의 발전을 저해했던 데이터 부족 문제를 어떻게 극복할 수 있는가? 둘째, 충분한 비디오 데이터를 사용할 수 있을 때 어떤 아키텍처 접근 방식이 가장 효과적인가? 그들의 해결책은 400가지 행동 클래스와 240,000개 이상의 훈련 비디오를 포함하는 Kinetics 데이터셋을 생성하고, 성공적인 2D 이미지 아키텍처를 시공간 학습에 영리하게 적용한 Inflated 3D ConvNets(I3D)를 개발하는 것을 포함합니다.

2. 데이터셋 기여: Kinetics 인간 행동 비디오 데이터셋

본 논문의 첫 번째 주요 기여는 비디오 이해 연구의 규모에 있어 패러다임의 전환을 나타내는 Kinetics 인간 행동 비디오 데이터셋입니다. 이 연구 이전에 가장 큰 행동 인식 데이터셋은 UCF-101 (101개 클래스에 걸쳐 13,320개 비디오)과 HMDB-51 (51개 클래스에 걸쳐 6,766개 비디오)이었습니다. 이러한 데이터셋은 초기 연구에는 유용했지만, 딥 뉴럴 네트워크를 효과적으로 훈련시키기에는 불충분했습니다.

Kinetics는 약 10초 길이의 비디오 클립을 각 400개 이상 포함하는 400가지 인간 행동 클래스를 제공함으로써 이러한 한계를 해결합니다. 전체 데이터셋은 약 240,000개의 훈련 비디오와 40,000개의 테스트 비디오로 구성되어 있으며, 기존 벤치마크에 비해 두 자릿수 규모 증가를 나타냅니다. 행동은 세 가지 범주로 나뉩니다: 사람 행동 (개별 활동), 사람-사람 행동 (사람 간의 상호작용), 사람-객체 행동 (객체와의 상호작용).

데이터셋의 비디오는 YouTube에서 가져왔으며, 시각적 외관, 카메라 시점, 비디오 품질 및 맥락적 배경의 다양성을 보장합니다. 이러한 실제 세계의 가변성은 Kinetics를 이전 데이터셋보다 훨씬 더 도전적으로 만듭니다. 이전 데이터셋은 종종 통제된 환경이나 제한된 맥락의 비디오를 포함했습니다. Kinetics의 규모와 다양성은 연구자들이 더 깊은 모델을 훈련하고 이전에 조사할 수 없었던 전이 학습 효과를 연구할 수 있게 합니다.

3. 아키텍처 혁신: Two-Stream Inflated 3D ConvNets

본 논문의 두 번째 주요 기여는 성공적인 2D 이미지 모델의 이점과 진정한 시공간 처리를 효과적으로 결합한 새로운 아키텍처인 Two-Stream Inflated 3D ConvNets(I3D)의 개발입니다. 핵심 통찰력은 3D 컨볼루션 네트워크가 비디오 처리에 자연스럽지만, 대규모 비디오 데이터셋의 부족과 사전 훈련된 이미지 모델을 활용할 수 없다는 점 때문에 효과적으로 훈련하기 어려웠다는 것입니다.

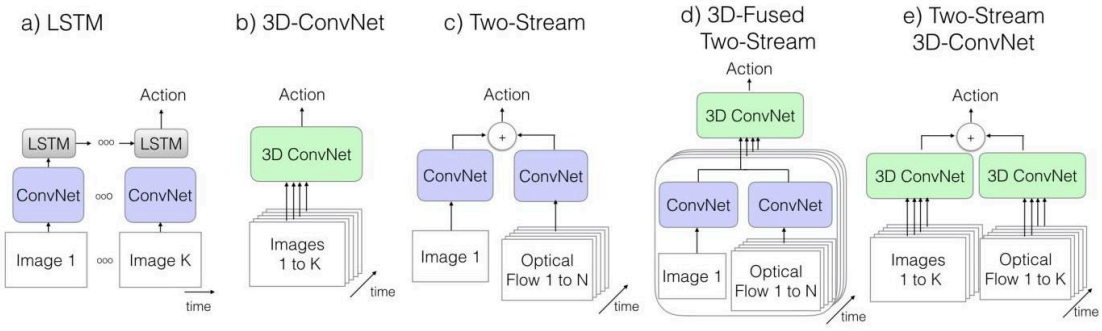

그림 2: 행동 인식에 대한 다섯 가지 다른 접근 방식 비교: (a) LSTM 기반 시간 모델링, (b) 3D ConvNets, (c) Two-Stream 네트워크, (d) 3D-융합 Two-Stream, (e) 제안된 Two-Stream I3D 아키텍처.

3.1. Inflation 과정

I3D의 핵심 혁신은 성공적인 2D 아키텍처를 3D 시공간 모델로 변환하는 "팽창(inflation)" 과정에 있습니다. 이 과정에는 몇 가지 주요 단계가 포함됩니다.

1. **필터 팽창**: N×N 크기의 2D 합성곱 필터가 N×N×N 크기의 3D 필터로 팽창되어, 움직임 패턴을 포착하기 위한 시간적 차원이 추가됩니다.

2. **파라미터 부트스트랩핑**: 저자들은 미리 훈련된 2D ImageNet 가중치를 사용하여 3D 필터를 초기화하는 원칙적인 방법을 개발했습니다. 그들은 "지루한 비디오 고정점(boring-video fixed point)" 원칙을 사용합니다. 즉, 정적 이미지가 반복되어 비디오를 형성하는 경우, 3D ConvNet은 원래 2D ConvNet과 동일한 출력을 생성해야 합니다. 이는 2D 필터 가중치를 시간 차원을 따라 N번 복제하고 N으로 나누어 달성됩니다.

$$w_{i,j,t} = \frac{w_{i,j}}{N}$$

여기서 $w_{i,j,t}$는 3D 필터 가중치이고 $w_{i,j}$는 원래 2D 가중치입니다.

3. **수용장 페이싱**: 저자들은 공간 및 시간 수용장 성장을 균형 있게 유지하기 위해 시간적 풀링을 신중하게 설계했습니다. 초기 레이어는 비대칭 풀링(1×3×3)을 사용하여 시간 해상도를 보존하는 반면, 후기 레이어는 대칭 풀링을 사용하여 더 넓은 시간 패턴을 포착합니다.

3.2. 투 스트림 구성

I3D가 원시 RGB 비디오를 직접 처리할 수 있음에도 불구하고, 저자들은 RGB 프레임과 광학 흐름(optical flow)을 별도로 처리하는 투 스트림(two-stream) 접근 방식이 우수한 성능을 제공한다는 것을 발견했습니다. 두 스트림은 다음과 같습니다.

- **RGB 스트림**: 원시 비디오 프레임을 처리하여 외관 정보를 캡처합니다.
- **흐름 스트림**: 미리 계산된 광학 흐름을 처리하여 움직임 정보를 캡처합니다.

최종 예측은 후기 융합(예측 평균)을 통해 두 스트림의 출력을 결합합니다.

4. 아키텍처 세부 정보 및 구현

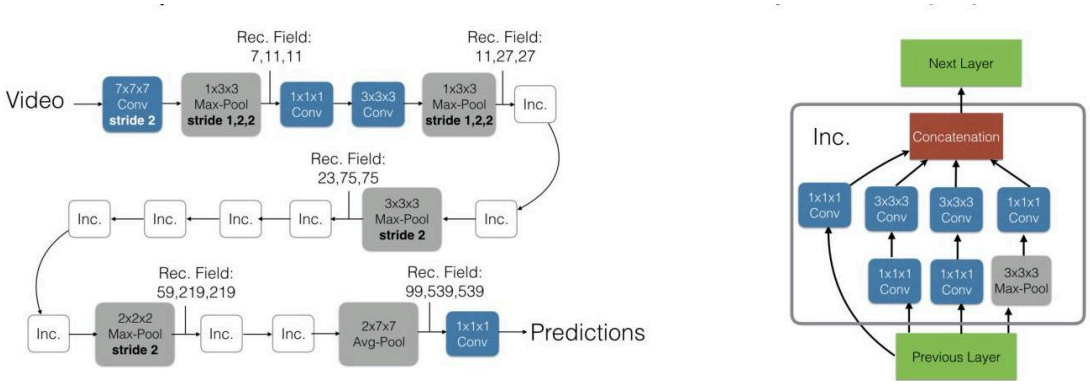

그림 3: Inflated 3D ConvNet의 상세 아키텍처로, Inception-v1 아키텍처가 3D 합성곱 및 풀링 연산을 통해 팽창되는 것을 보여줍니다. 이 다이어그램은 다른 레이어에서의 수용장 크기와 Inception 모듈 구조를 나타냅니다.

I3D 아키텍처는 ImageNet에서의 효율성과 강력한 성능으로 선택된 Inception-v1(GoogLeNet)을 기본 2D

아키텍처로 사용합니다. 팽창 과정은 모든 2D 연산을 해당 3D 연산으로 변환합니다.

- **3D 합성곱**: 모든 합성곱 레이어는 3D 커널을 사용하여 시공간 볼륨을 처리합니다.
- **3D 풀링**: 최대 풀링 연산은 3D로 확장되며, 시간적 스트라이드 선택에 신중을 기합니다.
- **Inception 모듈**: Inception 모듈의 복잡한 다중 브랜치 구조는 3D에서도 유지됩니다.

네트워크는 224×224 공간 해상도에서 64프레임의 비디오 클립을 처리합니다. 훈련 중에는 데이터 증강을 위해 무작위 시공간 자르기가 사용되며, 테스트 시에는 여러 클립을 샘플링하고 평균화하여 최종 예측을 수행합니다.

5. 실험 결과 및 검증

저자들은 액션 인식을 위한 다섯 가지 다른 접근 방식을 비교하는 포괄적인 실험을 수행했습니다.

1. ConvNet+LSTM
2. 3D ConvNet (C3D 스타일)
3. 투 스트림 네트워크
4. 3D 융합 투 스트림
5. 투 스트림 I3D (제안)

5.1. Kinetics 성능

Kinetics 데이터셋 자체에서 투 스트림 I3D가 최고의 성능을 달성했으며, 대규모 비디오 이해를 위한 팽창 접근 방식의 효율성을 입증했습니다. 결과는 충분한 훈련 데이터가 제공될 때 프레임 기반 접근 방식에 비해 시공간 처리의 명확한 이점을 보여줍니다.

5.2. 전이 학습 결과

가장 설득력 있는 결과는 UCF-101 및 HMDB-51에서의 전이 학습 실험에서 나옵니다. Kinetics에서 사전 훈련 후, 모든 아키텍처는 상당한 개선을 보였지만, 투 스트림 I3D는 놀라운 성능을 달성했습니다.

- **UCF-101**: 98.0% 정확도 (이전 최신 기술 대비 오류율 63% 감소)
- **HMDB-51**: 80.9% 정확도 (이전 최신 기술 대비 오류율 35% 감소)

이러한 결과는 대규모 비디오 사전 훈련이 이미지 작업에 대한 ImageNet 사전 훈련과 유사한 이점을 제공한다는 것을 보여줍니다.

6. 학습된 표현 분석

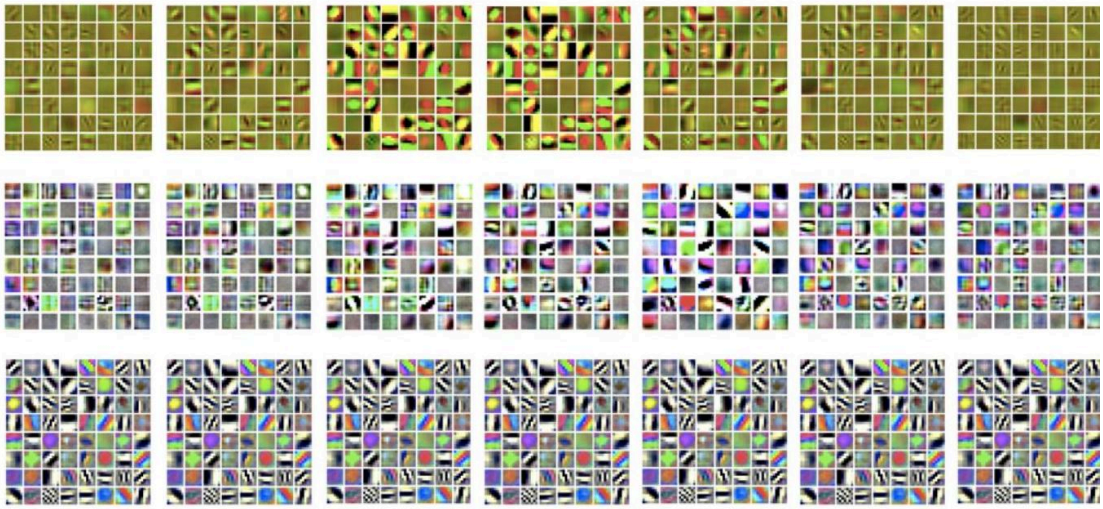

그림 4: 첫 번째 컨볼루션 레이어에서 학습된 3D 필터의 시각화. 2D ImageNet 필터(가장 왼쪽)에서 비디오 데이터 훈련 후 시간적으로 구조화된 3D 필터로 진화하는 모습을 보여줍니다. 각 행은 ImageNet 2D 필터, RGB I3D 필터, Flow I3D 필터 등 다른 단계를 나타냅니다.

이 논문은 필터 시각화를 통해 I3D 모델이 무엇을 학습하는지에 대한 귀중한 통찰력을 제공합니다. 분석 결과는 다음과 같습니다:

- **시간적 구조**: 2D ImageNet 필터와 달리 훈련된 I3D 필터는 풍부한 시간적 구조를 보여 모델이 움직임 패턴을 포착하는 방법을 학습했음을 나타냅니다.
- **스트림 차이**: RGB 및 플로우 스트림은 다른 유형의 필터를 학습하며, 플로우 스트림 필터가 원본 ImageNet 필터와 더 유사합니다.
- **특징 전이성**: Kinetics에서 학습된 특징은 최종 분류 레이어만 미세 조정했을 때 강력한 성능을 보이는 것으로 보아 다른 데이터셋으로도 잘 전이됩니다.

7. 영향 및 중요성

이 연구는 비디오 이해 연구에서 분수령이 되는 순간을 나타냅니다. 효과적인 시공간 학습에 필요한 데이터(Kinetics)와 아키텍처(I3D)를 모두 제공함으로써, 저자들은 비디오 분석을 위한 새로운 패러다임을 확립했습니다. 이 논문의 영향은 액션 인식뿐만 아니라 객체 탐지, 분할, 시간적 액션 위치 파악을 포함한 더 넓은 비디오 이해 작업에까지 미칩니다.

확장(inflation) 기술은 특히 영향력이 컸으며, 사전 훈련된 가중치를 활용하면서 성공적인 이미지 아키텍처를 비디오에 적용하는 원칙적인 방법을 제공했습니다. 이 접근 방식은 수많은 후속 연구에서 채택되고 확장되어 현대 비디오 이해의 근본적인 기술이 되었습니다.

8. 기술적 기여 및 방법론

이 논문은 몇 가지 주요 기술적 기여를 합니다:

1. **체계적인 아키텍처 비교**: 저자들은 대규모 데이터셋에서 액션 인식을 위한 다양한 딥러닝 접근 방식에 대한 최초의 포괄적인 비교를 제공하여, 데이터가 풍부할 때 어떤 아키텍처 선택이 가장 유익한지 밝혀냈습니다.
2. **확장(Inflation) 방법론**: 사전 훈련된 지식을 보존하면서 2D 아키텍처를 3D로 변환하는 원칙적인 접근 방식은 광범위하게 적용될 수 있는 중요한 방법론적 진보입니다.
3. **전이 학습 분석**: 이 연구는 전이 학습을 위한 비디오 기반 사전 훈련의 힘을 입증하고, 해당 분야의 표준이 된 모범 사례를 확립했습니다.

실험 방법론은 엄격하며, 구현 세부 사항, 아키텍처 간의 공정한 비교, 결과에 대한 철저한 분석에 세심한 주의를 기울였습니다. 저자들은 자신들의 모델을 공개적으로 이용 가능하게 하여 후속 연구를 촉진하고 재현성을 보장했습니다.

9. 결론

"액션 인식, 어디로 가는가?(Quo Vadis, Action Recognition?)"는 목적지(대규모 비디오 이해)와 그곳에 도달하기 위한 수단(I3D 아키텍처)을 모두 제공함으로써 제목의 질문에 성공적으로 답합니다. Kinetics 데이터셋과 Two-Stream I3D 아키텍처의 조합은 비디오 이해 연구를 근본적으로 변화시켰으며, 오늘날에도 여전히 이 분야에 영향을 미치는 성능과 방법론의 새로운 기준을 확립했습니다. 이 연구는 충분한 데이터와 적절한 아키텍처를 통해 비디오 이해가 이미지 인식에 혁명을 가져온 것과 동일한 극적인 발전을 이룰 수 있음을 보여주며, 컴퓨터 비전 전반에 걸쳐 더욱 정교한 비디오 분석 애플리케이션의 길을 열었습니다.

10. 관련 인용

동영상 내 동작 인식을 위한 이중 스트림 합성곱 신경망

본 논문은 외형과 움직임을 포착하기 위해 RGB 스트림과 광학 흐름 스트림을 별도로 처리하는 근간이 되는 투스트림 아키텍처를 소개합니다. 'Quo Vadis' 논문은 이를 비교를 위한 핵심 기준선으로 사용하며, 투스트림 개념을 자체 최고 성능의 I3D 모델에 통합합니다.

K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos.InAdvances in Neural Information Processing Systems, pages 568–576, 2014.

3차원 합성곱 신경망을 이용한 시공간 특징 학습

C3D로 알려진 이 연구는 본 논문이 비교 대상으로 삼는 3D ConvNet 아키텍처의 대표적인 예시입니다. 제안된 I3D 모델은 매우 깊은 2D 아키텍처를 활용하여 C3D와 같은 기존의 얕은 3D ConvNet의 한계를 극복하기 위한 방법으로 제시됩니다.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497. IEEE, 2015.

키네틱스 사람 동작 비디오 데이터셋

이 인용문은 본 논문의 핵심적인 기여이자 주요 실험 검증 기반인 Kinetics 데이터셋에 대해 설명합니다. 본 논문의 전체적인 논지는 이 새롭고 대규모 데이터셋을 사용하여 기존 모델들을 재평가하고, 사전 학습(pre-training)을 통해 얻는 상당한 성능 향상을 입증하는 데 있습니다.

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. The kinetics human action video dataset.arXiv preprint arXiv:1705.06950, 2017.

배치 정규화: 내부 공변량 변화 감소를 통한 딥 네트워크 훈련 가속화

저자들은 자신들의 새로운 Inflated 3D ConvNet (I3D) 모델이 배치 정규화를 포함한 Inception-v1 아키텍처를 기반으로 한다고 명시하며, 본 문헌이 이에 대한 핵심 참조 자료입니다. 본 인용은 새로운 모델을 생성하기 위해 3D로 "확장(inflated)"된 특정하고 성공적인 2D 아키텍처를 이해하는 데 매우 중요합니다.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.