# The “something something” video database for learning and evaluating visual common sense

https://arxiv.org/pdf/1706.04261

## 0. Introduction

- 기존의 이미지 분류 데이터셋(ImageNet 등)은 객체 인식에 초점을 맞추었으나 인간의 시각적 상식(visual common sense)을 반영한 동작 인식에는 한계가 있음
- 일상적인 동작을 이해하기 위해서는 무엇을 하는지에 대한 정교한 이해가 필요하며 이는 단순한 객체 분류로는 포착하기 어려움

## 1. Overview

- 시각적 상식(visual common sense)을 학습하고 평가하기 위한 새로운 비디오 데이터셋과 접근 방법 제안
- 비디오를 캡션 템플릿과 결합하여 "무엇을 하는지"에 대한 정교한 의미를 학습
- 174개의 동작 클래스 정의를 통해 다양한 행동 패턴 포착
- 캡션 템플릿 예시 : "Picking [something] up", "Throwing [something] away"

## 2. Challenges

- 데이터 특성 :
  - 무엇을 하는지에 대한 정교한 이해는 단순한 객체 분류로는 포착하기 어려움
- 크라우드소싱의 한계 :
  - 크라우드워커들이 캡션을 작성할 때의 일관성 부족

    - 다양한 해석이 가능하여 데이터 품질 관리 어려움
- 모델 학습의 어려움 :
    - 동작 인식 모델이 시각적 상식을 학습하기 위해서는 정교한 구조와 학습 전략 필요

# 3. Method

| Dataset Specifications | |
|---|---|
| Number of videos | 108,499 |
| Number of class labels | 174 |
| Average duration of videos (in seconds) | 4.03 |
| Average number of videos per class | 620 |

- 데이터 수집 :
    - 크라우드소싱 플랫폼을 통해 다양한 사람들이 일상적인 동작을 촬영한 비디오 수집
    - 각 비디오는 특정 동작을 설명하는 캡션 템플릿으로 라벨링됨
- 데이터 전처리 :
    - 비디오의 해상도와 프레임 속도를 통일하여 모델 학습에 적합하게 변환
- 모델 학습 :
    - 다양한 딥러닝 모델을 사용하여 동작 인식 성능 평가
    - 시각적 상식을 학습하기 위해 특별한 구조와 학습 전략 채택

# 4. Experiments

- 모델 비교 :
    - 기존 동작 인식 모델들과 본 데이터셋 사용 성능 비교
    - 다양한 모델이 시각적 상식을 학습하는 능력 평가
- 성능 평가 :
    - 정확도, 정밀도, 재현율 등의 지표 사용
    - 모델이 시각적 상식을 얼마나 잘 학습했는지 측정

# 5. Results

| Dataset | Domain | # Videos | Avg. duration | Remarks |
|---|---|---|---|---|
| Physics 101 [38] | intuitive physics | 17,408 | - | 101 objects with 4 different scenarios (ramp, spring, fall, liquid) |
| MPII cooking [26] | action (cooking) | 44 | 600s | - |
| TACoS [23] | action (cooking) | 127 | 360s | - |
| Charades [29] | action (human) | 10, 000 | 30s | - |
| KITTI [6] | action (driving) | 21 | 30s | - |
| Something-Something (ours) | human-object interaction | 108,499 | 4.03s | 174 fine-grained categories of human-object interaction scenarios |

| Method | Error rate (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 10 classes | | 40 classes | | 174 classes | | |
| | top-1 | top-2 | top-1 | top-2 | top-1 | top-2 | top-5 |
| 2D CNN + Avg | 76.5 | 58.9 | 88.0 | 78.5 | - | - | - |
| Pre-2D CNN + Avg | 54.7 | 39.0 | 79.2 | 70.0 | - | - | - |
| Pre-2D CNN + LSTM | 52.3 | 34.1 | 77.8 | 68.0 | - | - | - |
| 3D CNN + Stack | 58.1 | 38.7 | 70.3 | 57.3 | - | - | - |
| Pre-3D CNN + Avg | 47.5 | 29.2 | 66.2 | 52.7 | 88.5 | 81.5 | 70.0 |
| 2D+3D-CNN | **44.9** | **27.1** | **63.8** | **50.7** | - | - | - |

Table 4: Error rates on different subsets of the data.

- 본 데이터셋을 사용한 모델들이 기존 모델들보다 우수한 성능을 보임
- 특히 시각적 상식을 잘 학습한 모델들이 높은 정확도를 기록
- 다양한 모델들이 본 데이터셋에서 테스트되었으며 성능 차이 비교
- 일부 모델은 시각적 상식을 잘 학습하여 높은 성능 보임
- 본 데이터셋이 모델의 시각적 상식 학습에 중요한 역할
- 데이터셋의 다양성과 품질이 모델 성능에 큰 영향

# 6. Insight

- 동작 인식 모델이 인간처럼 동작을 이해하기 위해서는 시각적 상식 필수
- 데이터셋은 이러한 시각적 상식을 학습하는 중요한 자원
- 캡션 템플릿 사용으로 "무엇을 하는지" 명확하게 정의하여 모델이 시각적 상식을 학습하도록 유도