ACTIONS ~ TRANSFORMATIONS

Xiaolong Wang, Ali Farhadi, Abhinav Gupta
Carnegie Mellon University, University of Washington
The Allen Institute for AI
 ACTIONS ~ TRANSFORMATIONS

• Motivation
• Dataset
• Modeling Actions as Transformations
• Experiments
• Action Recognition

Aggregated Score

Score 1

Classify

Score 2

Classify

Score 3

Classify

Score 4

Classify
MOTIVATION

Precondition $\times$ Action $=$ Effect
**Motivation**

- Action: the transformation between precondition and effect.
  - Baseline: Direct classification on action
ACTION RECOGNITION

Training:

Testing:
**CROSS CATEGORY GENERALIZATION**

**Training:**

![Training images]

× Jump =

**Testing:**

![Testing images]
PREDICTION

\[ \times \text{Dive} = \]
**Actions ≈ Transformations**

- Motivation
- Dataset
- Modeling Actions as Transformations
- Experiments
DATASET

• We propose a new action dataset ACT, for testing cross-category generalization.

• 43 action classes, and 16 super classes.

• Youtube videos labeled by turkers: Around 11000 videos with 1-10 seconds.
DATASET

- 2-layer hierarchy

DATASET

- Catching: football, soccer;
- Climbing: rock, rope, tree;
- Crawling: baby, army;
- Cutting: apple, orange, pineapple, watermelon
- Diving: bunjeejumping, cliffdiving, divepool
- Jumping: highjump, longjump
- Kicking-fight: bag, people
- Kicking-ball: soccergoal, soccerpass, soccerpenalty
- Lifting: benchpress, cleanandjerk
- Pouring: beer, juice, milk
- Punching: bag, people, speedbag
- Pushing: car, cart
- Sliding: skateboard, ski, snowboard
- Swinging: baseball, golf, tennis
- Weaponfighting: fencing, swordfighting
- Throwing: discus, shotput, baseball, hammer
DATASET

• Task Design
  ▪ Standard Action Recognition (1 split).
  ▪ Cross-category: For each super class, leave one sub-category for testing and the others as training data (3 splits).
ACTIONS ≈ TRANSFORMATIONS

- Motivation
- Dataset
- Modeling Actions as Transformations
- Experiments
MODEL

• Input with latent variables

Precondition  Effect
MODEL

- Input samples

Precond frames

Effect frames

Network 1

Network 2
• Siamese Network (VGG 16)
TRAINING

• Learning Objective
  - \( \text{minimized} \) \( \text{dist}(\text{precond}\_\text{fc}\times T \downarrow y, \text{effect}\_\text{fc}) \)
  - \( \text{maximized} \) \( \text{dist}(\text{precond}\_\text{fc}\times T \downarrow i, \text{effect}\_\text{fc}) \) for \( i \neq y \)

• Iterate between 2 steps:
  - Learn model parameters
  - Estimate latent variables
TESTING

Precondition

Effect
TESTING

Precondition

Effect
TESTING

Precondition

Effect
ACTIONS ≈ TRANSFORMATIONS

• Motivation
• Dataset
• Modeling Actions as Transformations
• Experiments
NUMBERS

- UCF101: 13000 videos, 101 classes

<table>
<thead>
<tr>
<th>Method</th>
<th>RGB</th>
<th>Optical Flow</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Stream [32]</td>
<td>73.0%</td>
<td>83.7%</td>
<td>88.0%</td>
</tr>
<tr>
<td>Two Stream (VGG16) [45]</td>
<td>78.4%</td>
<td>87.0%</td>
<td>91.4%</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>80.8%</strong></td>
<td><strong>87.8%</strong></td>
<td><strong>92.4%</strong></td>
</tr>
</tbody>
</table>

- HMDB51: 6700 videos, 51 classes

<table>
<thead>
<tr>
<th>Method</th>
<th>RGB</th>
<th>Optical Flow</th>
<th>Ave Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Stream [32]</td>
<td>40.5%</td>
<td>54.6%</td>
<td>58.0%</td>
</tr>
<tr>
<td>Two Stream (VGG16)</td>
<td>42.2%</td>
<td>55.0%</td>
<td>58.5%</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>44.1%</strong></td>
<td><strong>57.1%</strong></td>
<td><strong>62.0%</strong></td>
</tr>
</tbody>
</table>
NUMBERS

• Training precondition and effect models separately (UCF)

<table>
<thead>
<tr>
<th>Method (for split1)</th>
<th>RGB</th>
<th>Optical Flow</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Stream (VGG16) [45]</td>
<td>79.8%</td>
<td>85.7%</td>
<td>90.9%</td>
</tr>
<tr>
<td>Two Stream First Half</td>
<td>80.0%</td>
<td>84.7%</td>
<td>90.0%</td>
</tr>
<tr>
<td>Two Stream Second Half</td>
<td>79.5%</td>
<td>84.8%</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>81.9%</td>
<td>86.4%</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

• Ensemble with 2-stream (HMDB)
  ▪ 62%→63.4%
- ACT dataset, cross category generalization

<table>
<thead>
<tr>
<th>Model</th>
<th>RGB</th>
<th>Optical Flow</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Split1</td>
<td>Split2</td>
<td>Split3</td>
</tr>
<tr>
<td>Two Stream</td>
<td>48.3%</td>
<td>53.2%</td>
<td>49.8%</td>
</tr>
<tr>
<td>LSTM+Two Stream</td>
<td>47.7%</td>
<td>53.4%</td>
<td>50.4%</td>
</tr>
<tr>
<td>Ours</td>
<td>50.7%</td>
<td>55.2%</td>
<td>52.3%</td>
</tr>
</tbody>
</table>

**NUMBERS**
VISUALIZATION

• Back-prop gradients
Nearest Neighbor Testing

Query

Ours

Two Stream
Nearest Neighbor Testing

Query    Ours    Two Stream
Nearest Neighbor Testing

Query

Ours

Two Stream
PREDICTION RESULTS

Precondition Query

Effect Retrieval (same class)

Effect Retrieval (different class)
PREDICTION RESULTS

Precondition Query

Effect Retrieval (same class)

Effect Retrieval (different class)