Human action recognition and the Kinetics dataset

Andrew Zisserman

Includes slides from Joao Carreira and Rohit Girdhar
Outline

1. The Kinetics human action video dataset

2. Action recognition by pre-training on Kinetics

3. Where next in action recognition?
The Kinetics Human Action Video Dataset

archery  country line dancing  riding or walking with horse  playing violin  eating watermelon
Motivation

Objective: A large scale human action classification video dataset

- An ImageNet for human action recognition
  - Trimmed videos
  - Actions performed by humans
  - Action classification

- Large enough to use for architecture design and comparison

- Large enough to pre-train networks for other tasks, e.g.
  - Temporal action localization in untrimmed videos
Kinetics overview

- Stats:

<table>
<thead>
<tr>
<th></th>
<th>Year</th>
<th>Actions</th>
<th>Clips per class</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinetics-400</td>
<td>2017</td>
<td>400</td>
<td>400-1000</td>
<td>300k</td>
</tr>
<tr>
<td>Kinetics-600</td>
<td>2018</td>
<td>600</td>
<td>600-1000</td>
<td>500k</td>
</tr>
</tbody>
</table>

- 10s clips

- Every clip is from a different YouTube video
  - For each action, huge variety in people, viewpoint, execution ...

- A Short Note about Kinetics-600, Carreira, Noland, Bank-Horvath, Hillier, Zisserman, arXiv 2018
Action Classes

**Person Actions (Singular)**
e.g. waving, blinking, running, jumping

**Person-Person Actions**
e.g. hugging, kissing, shaking hands

**Person-Object Actions**
e.g. opening door, mowing lawn, washing dishes
Person Actions (Singular)

- Pumping Fist
- Shaking Head
Person Actions (Singular)

Long Jump

Triple Jump
Person-Person Actions

Shaking Hands

Massaging Back
Person-Object Actions

Playing Violin

Playing Trumpet
Person-Object Actions

Folding Clothes

Folding Napkin
Person-Object Actions

- **Planting Flowers**
- **Arranging Flowers**
Dataset Collection Pipeline

Class list

- 0 abseiling
- 1 laughing
- 2 swimming
- 3 shearing sheep
- 4 motorcycling
- 5 celebrating
- 6 spray painting
- 7 playing tennis
- 8 driving tractor
- 9 washing dishes
- 10 skateboarding
- 11 waxing legs

YouTube querying

“Playing drums”

Image Classifiers

Human verification using Mechanical Turk

Evaluating Actions in Videos

Does this video clip contain the human action playing drums?

- Yes, this is a true example of the action
- No, this is not an example of the action
- You are unsure if this is an example of the action
- Replay the video

Combine, split, and filter classes
Scaling up from 400x400 to 600x600

• Finding candidate videos
  • Kinetics-400: text query for class name
  • Kinetics-600: decouple class and query text, add concept of language

• e.g: "folding paper" now matches against
  • "folding paper" (en)
  • "origami" (en)
  • "dobrar papel" (pt)
Dataset Collection Pipeline

YouTube querying

“Drumming”
“Playing drums”
“Tocar bateria”

Image Classifiers

Human verification using Mechanical Turk

Combine, split, and filter classes
New in Kinetics-600: more body-only classes

Head stand

Tiptoeing
More face classes

Raising eyebrows

Crossing eyes
More hand classes

Twiddling fingers

Cracking knuckles
More basic tool use

Using sledgehammer

Using power drill

Also using paint roller, circular saw, wrench, others
More actions around similar objects

- Popping balloons
- Inflating balloons
- Throwing water balloons
- Making balloon shapes
More dances

Mosh pit dancing

Square dancing
More random stuff many people do

Contact juggling

Alligator wrestling
Comparison of networks on Kinetics

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td># Input Frames</td>
<td>Temporal Footprint</td>
</tr>
<tr>
<td>ConvNet+LSTM</td>
<td>9M</td>
<td>25 rgb</td>
<td>5s</td>
</tr>
<tr>
<td>3D-ConvNet</td>
<td>79M</td>
<td>16 rgb</td>
<td>0.64s</td>
</tr>
<tr>
<td>Two-Stream</td>
<td>12M</td>
<td>1 rgb, 10 flow</td>
<td>0.4s</td>
</tr>
<tr>
<td>3D-Fused</td>
<td>39M</td>
<td>5 rgb, 50 flow</td>
<td>2s</td>
</tr>
<tr>
<td>Two-Stream 3D</td>
<td>25M</td>
<td>64 rgb, 64 flow</td>
<td>2.56s</td>
</tr>
</tbody>
</table>

Table 1. Number of parameters and temporal input sizes of the models.

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
Joao Carreira, Andrew Zisserman, CVPR 17
Inflated 3D Inception (I3D)

Inflated Inception-V1

Inception Module (Inc.)

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
Joao Carreira, Andrew Zisserman, CVPR 17
## Network comparison on Kinetics-400

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Kinetics</th>
<th>ImageNet then Kinetics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RGB</td>
<td>Flow</td>
</tr>
<tr>
<td>(a) LSTM</td>
<td>53.9</td>
<td>-</td>
</tr>
<tr>
<td>(b) 3D-ConvNet</td>
<td>56.1</td>
<td>-</td>
</tr>
<tr>
<td>(c) Two-Stream</td>
<td>57.9</td>
<td>49.6</td>
</tr>
<tr>
<td>(d) 3D-Fused</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(e) Two-Stream I3D</td>
<td><strong>68.4 (88.0)</strong></td>
<td><strong>61.5 (83.4)</strong></td>
</tr>
</tbody>
</table>

Table 3. Performance training and testing on Kinetics with and without ImageNet pretraining. Numbers in brackets () are the Top-5 accuracy, all others are Top-1.

---

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset  
Joao Carreira, Andrew Zisserman, CVPR 17
I3D comparison from Kinetics-400 to Kinetics-600

<table>
<thead>
<tr>
<th>Model</th>
<th>ImageNet + Kinetics</th>
<th>Kinetics</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB-I3D,</td>
<td>71.1 / 89.3</td>
<td>68.4 / 88.0</td>
</tr>
<tr>
<td>Flow-I3D,</td>
<td>63.4 / 84.9</td>
<td>61.5 / 83.4</td>
</tr>
<tr>
<td>Two-Stream I3D</td>
<td>74.2 / 91.3</td>
<td>71.6 / 90.0</td>
</tr>
</tbody>
</table>

Kinetics-600, RGB-I3D, training/testing on Kinetics-600: 72.0 / 91.0

A Short Note about Kinetics-600
Authors: Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, Andrew Zisserman, arXiv 2018
Part II

Action recognition by pre-training on Kinetics

Performance on four datasets:
1. UCF-101 – classification
2. HMD-51 – classification
3. Charades – temporal localization
4. AVA – spatio-temporal localization
UCF-101 and HMDB-51

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Actions</th>
<th>Clips</th>
<th>Total</th>
<th>Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMDB-51 [15]</td>
<td>2011</td>
<td>51</td>
<td>min 102</td>
<td>6,766</td>
<td>3,312</td>
</tr>
<tr>
<td>UCF-101 [20]</td>
<td>2012</td>
<td>101</td>
<td>min 101</td>
<td>13,320</td>
<td>2,500</td>
</tr>
</tbody>
</table>
Transferring from ImageNet to Video

UCF-101

Best method using just hand designed features

HMDB-51

Best method using just hand designed features

Compilation of results from actionrecognition.net
I3D-Kinetics-400 transfer performance (two stream, flow+RGB)

Compilation of results from actionrecognition.net
Charades dataset - action localization

- I3D model with Kinetics-400 pre-training defined the state of the art
- Winner of the CVPR 2017 Charades challenge
Atomic Visual Actions (AVA) Dataset

- Person-centric actions
- Multiple people, multiple action labels
- Atomic actions
- Exhaustivity
- Action transitions over time
- Realistic scenes and diverse environment

## 80 Atomic Actions in AVA

<table>
<thead>
<tr>
<th>Group</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pose (14)</td>
<td>run/jog, walk, jump, stand, sit, lie/sleep, bend/bow, crawl, swim, dance, get up, fall down, crouch/kneel, martial art</td>
</tr>
<tr>
<td>Person-Person (17)</td>
<td>talk to, watch, listen to, sing to, kiss, hug, grab, lift, kick, give/serve to, take from, play with kids, hand shake, hand clap, hand wave, fight/hit, push</td>
</tr>
<tr>
<td>Person-Object (49)</td>
<td>lift/pick up, put down, carry, hold, throw, catch, eat, drink, cut, hit, stir, press, extract, read, write, smoke, sail boat, row boat, fishing, touch, cook, paint, dig, shovel, chop, shoot, take a photo, brush teeth, clink glass, work on a computer, answer phone, climb (e.g., mountain), play board game, play with pets, drive (e.g., a car), push (an object), pull (an object), point to (an object), open, close, enter, exit, play musical instrument, text on/look at a cellphone, turn (e.g., screwdriver), dress / put on clothing, ride (e.g., bike, car, horse), watch (e.g., TV)</td>
</tr>
</tbody>
</table>
AVA Challenge 2018

Localize the atomic actions in space & time

Frame mAP @ >0.5 IoU

on 1 fps keyframes of 15-minute segments
from 131 test videos
Model overview

*Image Diagram*:

```
T x H x W x 3
RGB frames

RGB
I3D

ROI Pooling

Avg Pooling

H' x W' x C

Classification
Box Refinement
```

*References*:

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018
Network architecture

1. Extract clip-features using I3D

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018
Network architecture

1. Extract clip-features using I3D
2. Compute regions on center frame features
3. Extend regions temporally

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018
Network architecture

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018
Groundtruth

- watch (a person) (50,68,25,97)
- listen to (a person) (62,75,39,99)
- watch (a person) (62,75,39,99)
- **grab (a person) (50,68,25,97)**
- bend/bow (at the waist) (50,68,25,97)
- watch (a person) (35,51,56,99)
- listen to (a person) (35,51,56,99)
- stand (35,51,56,99)
- stand (62,75,39,99)

*A Better Baseline for AVA,*
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018
Predictions

Test set mAP = 21%

A Better Baseline for AVA,
Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, arXiv 2018
Easiest and Hardest Classes

Bars color coded by dataset size of the class. Lighter colors are higher.
Part III

Where next in action recognition?
Video

A temporal sequence of frames

What is required to recognize the action?

• a single frame?

• a bag of frames (unordered)?

• an ordered sequence of frames?

• ...
Action Classification on Static Frames

- Jumping
- Phoning
- Playing Instrument
- Reading
- Riding Bike
- Riding Horse
- Running
- Taking Photo
- Using Computer
- Walking

PASCAL VOC Action Classification Challenge
Some actions require motion for classification

• Sitting down/standing up; closing/opening something
• Different dance styles ....
Some actions require motion for classification

• Sitting down/standing up; closing/opening something
• Different dance styles ....

Dancing Macarena  Dancing Charleston  Zumba
Representing motion using optical flow

- Throws away “nuisance factors” like appearance of clothes and skin
- Helps with foreground/background segmentation
The benefits of optical flow

- **Two-Stream ConvNet Architecture**
  - Appearance stream ConvNet (input: still RGB frames)
  - Temporal stream ConvNet (input: multi-frame optical flow)

- **UCF-101 Mean Accuracy (across all splits)**
  
<table>
<thead>
<tr>
<th>Model</th>
<th>UCF-101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Stream ConvNet</td>
<td>72.6</td>
</tr>
<tr>
<td>Temporal Stream ConvNet (multi-task)</td>
<td>83.6</td>
</tr>
<tr>
<td>Two-stream fusion (by averaging)</td>
<td>86.9</td>
</tr>
<tr>
<td>Two-stream fusion (weighted averaging)</td>
<td>87.6</td>
</tr>
</tbody>
</table>

*K. Simonyan, A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", NIPS 2014*
State of the art on Kinetics-400

Top-1 % accuracy on Action classification performance on Kinetics-400 val

<table>
<thead>
<tr>
<th>Model</th>
<th>RGB only</th>
<th>RGB + flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3D-G</td>
<td>74.7</td>
<td>77.2</td>
</tr>
<tr>
<td>TSN Inception V3</td>
<td>72.5</td>
<td>76.6</td>
</tr>
<tr>
<td>Non-local Neural Networks</td>
<td>77.7</td>
<td></td>
</tr>
<tr>
<td>I3D</td>
<td>71.1</td>
<td>74.2</td>
</tr>
</tbody>
</table>

- Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification, Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, Kevin Murphy, ECCV 2018
- Non-local Neural Networks, Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, CVPR 2018
- Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Joao Carreira, Andrew Zisserman, CVPR 17
State of the art on Kinetics-400

Top-1 % accuracy on Action classification performance on Kinetics-400 val

<table>
<thead>
<tr>
<th>Model</th>
<th>RGB only</th>
<th>RGB + flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3D-G</td>
<td>74.7</td>
<td>77.2</td>
</tr>
<tr>
<td>TSN Inception V3</td>
<td>72.5</td>
<td>76.6</td>
</tr>
<tr>
<td>Non-local Neural Networks</td>
<td>77.7</td>
<td></td>
</tr>
<tr>
<td>I3D</td>
<td>71.1</td>
<td>74.2</td>
</tr>
</tbody>
</table>

- Ceiling on performance is currently less than 80%
- Adding flow boosts performance by around 3%
- Conclusion: RGB models are not able to fully learn from the motion information yet
What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets
De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles, CVPR 2018

• Conclusion: the C3D model (using 16 frames) does not use motion to classify 35% of the classes in Kinetics-400

• Consequently: either the model can not learn from the motion of those classes, or the classes do not require motion to classify them
Summary

Current generation of neural network architectures for action classification
  • Have not saturated performance on Kinetics yet
  • Are probably not learning motion information to its full potential

• Need for more innovation ... research questions:
  • How to develop architectures that can efficiently learn motion information?
  • How to develop lighter architectures for action classification?

Notes for the future:
  • Kinetics-800 will be released next year
  • ActivityNet workshop for Kinetics and AVA challenges