Transferring from Kinetics

João Carreira

Tutorial on Action Classification and Video Modelling
CVPR 2019
16th of June
Brief Recent History of Image Understanding

Caltech 101 (10k images):
- classification

2001-2012

Simple features

Problems studied in isolation

PASCAL VOC (10k images):
- object localization
Brief Recent History of Image Understanding

ImageNet (1M images):
- classification

2012-now

Deep learning
Transfer learning

PASCAL VOC / COCO:
- object localization
Finetuning ImageNet models on other classification datasets (2013)

DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition (Donahue et al)
Brief Recent History of Video Understanding

UCF 101 (10k videos) / HMDB-51 (5k videos):
- classification

2012-2016
ActivityNet, Thumos, UCF101-Det:
- Action localization

Problems studied in isolation
Transfer from ImageNet

Diving, Golf Swing, Kicking,
Running, Skateboarding, Swing-Bench
Transferring from ImageNet to Video

Compilation of results from actionrecognition.net

UCF-101

HMDB-51

Best method using just hand designed features
Ideal: learn representations directly from videos

Capture motion

Gunnar Johannson, video from 1971
Ideal: learn representations directly from videos

Image architectures wasteful for processing high-frame rate video
Deep learning on videos

Kinetics-400 (300k videos) - classification

2017

ActivityNet, Charades, AVA - Action localization

Deep learning video models on Kinetics-400

Transfer from Kinetics-400
1. The Kinetics dataset

- archery
- country line dancing
- riding or walking with horse
- playing violin
- eating watermelon
Kinetics-400 (2017)

- Object classification
- Human action classification (10s clips)

- 1000 object classes x 1000 images
- 400 human action classes x >400 videos
- (300k total, ~all from unique videos)

Images from google searches
Videos from youtube searches
Previous human action classification datasets too tiny to properly research new video representations

<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>
<tr>
<td>Kinetics</td>
<td>2017</td>
<td>400</td>
<td>min 400</td>
<td>306,245</td>
<td>306,245</td>
</tr>
</tbody>
</table>
Dataset Collection

Title matching

How to make healthy eating unbelievably easy | Luke TEDx Talks

Image Classifiers

Human verification using Mechanical Turk

Evaluating Actions in Videos

Instructions
We would like to find videos that contain real humans performing actions, e.g., sciencing their face, jumping, losing someone etc.

Please click on the most appropriate button after watching each video:

- 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

Images does not play, does not contain a human, is an image, cartoon or a computer game.

Combine, split, and filter classes
Action list

**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
Action list

Person Actions (Singular)

Pumping Fist

Shaking Head
Action list

Person Actions (Singular)

- Long Jump
- Triple Jump
Action list

Person-person actions

Shaking Hands

Massaging Back
Action list

Person-object actions

Playing Violin

Playing Trumpet
Action list

Person-object actions

Folding Clothes
Folding Napkin
Action list

Person-object actions

Planting Flowers

Arranging Flowers
DeepMind Shows AI Has Trouble Seeing Homer Simpson's Actions

The best artificial intelligence still has trouble visually recognizing people performing many of Homer Simpson's favorite behaviors such as drinking beer, eating chips, eating doughnuts, yawning, and the occasional face-plant. Those findings from DeepMind, the pioneering London-based AI lab, also suggest the motive behind why DeepMind has created a huge new dataset of YouTube clips to help train AI on identifying human actions in videos that go well beyond "Mmm, doughnuts" or "Doh!"

The most popular AI used by Google, Facebook, Amazon, and other companies beyond Silicon Valley is based on deep learning algorithms that can learn to identify patterns in huge amounts of data. Over time, such
Kinetics has kept growing

<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>
<tr>
<td>Kinetics</td>
<td>2017</td>
<td>400</td>
<td>min 400</td>
<td>306,245</td>
<td>306,245</td>
</tr>
</tbody>
</table>

<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>Kinetics-600</td>
<td>2018</td>
<td>600</td>
<td>min 450</td>
<td>500,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Kinetics-700</td>
<td>2019</td>
<td>700</td>
<td>min 450</td>
<td>650,000</td>
<td>650,000</td>
</tr>
</tbody>
</table>
Kinetics has kept growing

Looking in Mirror

Shoot dance
Other candidates to fill in for ImageNet for action recognition

- Sports-1M: 478 sports classes
- Something-Something: 174 classes, scripted
- Moments in Time: 339 “verb” classes (not just human)
- HACS: 200 classes + positive/negative samples
2. Transferring from Kinetics

Kinetics-400 (300k videos) - classification

2017

ActivityNet, Charades, AVA - Action localization

Deep learning video models on Kinetics-400

Transfer from Kinetics-400
Quo Vadis, Action Recognition?
A New Model and the Kinetics Dataset

- Comparison of models

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th># Input Frames</th>
<th>Temporal Footprint</th>
<th># Input Frames</th>
<th>Temporal Footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvNet+LSTM</td>
<td>9M</td>
<td>25 rgb</td>
<td>5s</td>
<td>50 rgb</td>
<td>10s</td>
</tr>
<tr>
<td>3D-ConvNet</td>
<td>79M</td>
<td>16 rgb</td>
<td>0.64s</td>
<td>240 rgb</td>
<td>9.6s</td>
</tr>
<tr>
<td>Two-Stream</td>
<td>12M</td>
<td>1 rgb, 10 flow</td>
<td>0.4s</td>
<td>25 rgb, 250 flow</td>
<td>10s</td>
</tr>
<tr>
<td>3D-Fused</td>
<td>39M</td>
<td>5 rgb, 50 flow</td>
<td>2s</td>
<td>25 rgb, 250 flow</td>
<td>10s</td>
</tr>
<tr>
<td>Two-Stream 3D</td>
<td>25M</td>
<td>64 rgb, 6 flow</td>
<td>2.56s</td>
<td>250 rgb, 250 flow</td>
<td>10s</td>
</tr>
</tbody>
</table>

Table 1. Number of parameters and temporal input sizes of the models.
Video-specific representations considered: 3D ConvNets

Example architecture: C3D

Learning Spatiotemporal Features with 3D Convolutional Networks. Tran et al, CVPR 2015

Figure 3. C3D architecture. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.
Video-specific representations considered: 3D ConvNets

Example architecture: C3D

Pure video model, that learns a hierarchical representation directly over video

The catch back then: performance was lower than two-stream networks. (e.g. UCF101):

<table>
<thead>
<tr>
<th>Model Description</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3D (1 net) + linear SVM</td>
<td>82.3</td>
</tr>
<tr>
<td>C3D (3 nets) + linear SVM</td>
<td>85.2</td>
</tr>
<tr>
<td>Two-stream networks [36]</td>
<td>88.0</td>
</tr>
</tbody>
</table>
Google’s Inception-V1 ImageNet classifier

*Going deeper with convolutions*, Szegedy et al, CVPR 2015
Inflated 3D Inception (I3D)
I3D Conv1 filters, trained in Kinetics

RGB

Flow
Transfer results with miniKinetics pre-training (80k videos)

Table 3. Performance on the UCF-101 and HMDB-51 test sets (splits 1 of both) for architectures pre-trained on miniKinetics. All except 3D-ConvNet are based on Inception-v1 and start off pre-trained on ImageNet. Original: train on UCF-101 / HMDB-51; Fixed: features from miniKinetics, with the last layer trained on UCF-101 / HMDB-51; Full-FT: miniKinetics pre-training with end-to-end fine-tuning on UCF-101 / HMDB-51; Δ shows the difference in misclassification as percentage between Original and the best of Full-FT and Fixed.
Comparison with state-of-the-art
Comparison with state-of-the-art

UCF-101 Test Set, Error (%)

- Previous SOTA (Feichtenhofer et al. 2016)
- I3D (UCF-101 only)
- I3D (UCF-101 + Kinetics)

HMDB-51 Test Set, Error (%)

- Previous SOTA (Feichtenhofer et al. 2016)
- I3D (HMDB-51 only)
- I3D (HMDB-51 + Kinetics)
Comparison with state-of-the-art
Performance as function of 
# Kinetics examples

![Graph showing performance as a function of kinetic examples]

- UCF-101
- HMDB-51
I3D-Kinetics-400 transfer performance (two stream, flow+rgb)

UCF-101

HMDB-51

Kinetics pre-training, comparison with state-of-the-art (compilation of results from actionrecognition.net)
Charades challenge winning entry at CVPR 2017

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Accuracy (mAP)</th>
<th>Modeling Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TeamKinetics</td>
<td>0.3441</td>
<td>I3D ConvNet with dense per-frame outputs</td>
</tr>
<tr>
<td>2</td>
<td>DR/OBU</td>
<td>0.2974</td>
<td>Two parallel convolutional neural networks (CNNs) extracting static (i.e., independent) appearance and optical flow features and scores for each frame, plus, there is another parallel audio feature extraction stream using Soundnet CNN, which is scored using a SVM.</td>
</tr>
<tr>
<td>3</td>
<td>UMICHL-VL</td>
<td>0.2811</td>
<td>We build an ensemble of Temporal Hourglass Networks (THGs), a novel architecture which consists of temporal convolutional layers, applied to several types of frame-wise feature vectors.</td>
</tr>
</tbody>
</table>
Charades challenge winning entry at CVPR 2017

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Accuracy (mAP)</th>
<th>Modeling Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TeamKinetics</td>
<td>0.2072</td>
<td>I3D ConvNet with dense per-frame outputs</td>
</tr>
<tr>
<td>2</td>
<td>UMich-VL</td>
<td>0.1803</td>
<td>We build an ensemble of Temporal Hourglass Networks (THGs), a novel architecture which consists of temporal convolutional layers, applied to several types of frame-wise feature vectors.</td>
</tr>
<tr>
<td>3</td>
<td>DR/OBU</td>
<td>0.1796</td>
<td>Two parallel convolutional neural networks (CNNs) extracting static (i.e., independent) appearance and optical flow features and scores for each frame, plus, there is another parallel audio feature extraction stream using Soundnet CNN, which is scored using a SVM.</td>
</tr>
</tbody>
</table>
Charades dataset

Video, 224x224 center crop

Top 5 + g.t. predictions

1. Holding a bag
2. Snuggling with a blanket
3. Lying on a bed
4. Opening a bag
5. Holding a blanket
6. Putting a bag somewhere
7. Someone is going from standing to sitting
8. Putting groceries somewhere
9. Sitting in a chair
10. Putting something on a table
11. Taking food from somewhere
12. Putting some food somewhere
13. Holding some medicine
Publications


2. *Quo Vadis Action Recognition: a New Model and the Kinetics Dataset*. Carreira and Zisserman, CVPR 2017
Conclusions

- Strengths:
  - Pretraining on Kinetics seems generally helpful
  - 3D ConvNets perform and transfer well

- Weaknesses:
  - Does not cover mid and long-term temporal modelling
  - Not appropriate directly as a curriculum for deployable robots to learn about human actions
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions

Chunhui Gu, Chen Sun, David Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik from Google Research

June 20, 2018 at Salt Lake City, CVPR18
Why a New Action Dataset?

- Person-centric actions
- Atomic actions
- Multiple actions over single person
- Exhaustivity
- Action transitions over time
- Realistic scenes and diverse environment
AVA Examples: Answer Phone
AVA Examples: Clink Glass
AVA Examples: Dig
AVA Examples: Give/Serve (object) to (person)
80 Atomic Actions in AVA

- run/jog
- walk
- jump
- stand
- sit
- lie/sleep
- bend/bow
- crawl
- swim
- dance
- get up
- fall down
- crouch/kneel
- martial art

Pose (14)
- 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

Person-Person (17)
- lift/pick up
- put down
- carry
- hold
- throw
- catch
- eat
- drink
- cut
- hit
- stir
- press
- extract
- read
- write

Person-Object (49)
- smoke
- sail boat
- row boat
- fishing
- touch
- cook
- kick
- paint
- dig
- shovel
- chop
- shoot
- take a photo
- brush teeth
- clink glass

Person-Object (49)
- 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)
- 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)

open
close
enter
exit
Atomicity from 3-sec segment sampled at 1Hz

Left: Kneel, Talk to
Right: Stand, Listen, Shoot
Pipeline Overview

YouTube Movie Selection ➔ Person Box Annotation ➔ Action Annotation ➔ Person Linking

3s

sit, ride, read
Dataset Statistics
AVA Dataset Size

- Number of videos: 430
- Number of segments: 386K
- Number of labeled bounding boxes: 614K
- Number of person tracks: 81K
- Number of labeled actions: 1.58M
Label Frequency

Long-tail distribution of action classes
## Action Transition over Time

<table>
<thead>
<tr>
<th>First Action</th>
<th>Second Action</th>
<th>NPMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watch (TV/monitor)</td>
<td>Work on a computer</td>
<td>0.64</td>
</tr>
<tr>
<td>Open (window/door)</td>
<td>Close (door/box)</td>
<td>0.59</td>
</tr>
<tr>
<td>Text on/Look at a cell phone</td>
<td>Answer phone</td>
<td>0.53</td>
</tr>
<tr>
<td>Listen to (a person)</td>
<td>Talk to (a person)</td>
<td>0.47</td>
</tr>
<tr>
<td>Fall down</td>
<td>Lie/Sleep</td>
<td>0.46</td>
</tr>
<tr>
<td>Talk to (a person)</td>
<td>Listen to (a person)</td>
<td>0.43</td>
</tr>
<tr>
<td>Stand</td>
<td>Sit</td>
<td>0.40</td>
</tr>
<tr>
<td>Walk</td>
<td>Stand</td>
<td>0.40</td>
</tr>
</tbody>
</table>
# Action Co-occurrence among Persons

<table>
<thead>
<tr>
<th>Person 1 Action</th>
<th>Person 2 Action</th>
<th>NPMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ride (bike/car/horse)</td>
<td>Drive (car/truck)</td>
<td>0.60</td>
</tr>
<tr>
<td>Play musical instrument</td>
<td>Listen to (music)</td>
<td>0.57</td>
</tr>
<tr>
<td>Take (object)</td>
<td>Give/Serve (object)</td>
<td>0.51</td>
</tr>
<tr>
<td>Talk to (a person)</td>
<td>Listen to (a person)</td>
<td>0.46</td>
</tr>
<tr>
<td>Stand</td>
<td>Sit</td>
<td>0.31</td>
</tr>
<tr>
<td>Play musical instrument</td>
<td>Dance</td>
<td>0.23</td>
</tr>
<tr>
<td>Watch (a person)</td>
<td>Write</td>
<td>0.15</td>
</tr>
<tr>
<td>Walk</td>
<td>Run/Jog</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Baseline Performance
Original Baseline 1

Faster R-CNN with ResNet-101 from ImageNet

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>11.3</td>
</tr>
</tbody>
</table>
Original AVA model – Baseline 2
Original Baseline 2

Flow I3D from Kinetics-400 + RGB I3D from Kinetics-400 + ResNet-50 from ImageNet (in Faster R-CNN framework)

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>11.3</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>15.6</td>
</tr>
</tbody>
</table>
AVA challenge 2018:

RGB I3D (in Faster R-CNN framework)

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>11.3</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>15.6</td>
</tr>
<tr>
<td>Ours</td>
<td>21.0</td>
</tr>
</tbody>
</table>

Other key differences:
- Data augmentation
- Class-agnostic bounding box regressor
AVA challenge 2018

Task #1 - Computer Vision

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Username</th>
<th>Organization</th>
<th><a href="mailto:mAP@0.5IoU">mAP@0.5IoU</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jianwen Jiang</td>
<td>Tsinghua University</td>
<td>21.08</td>
</tr>
<tr>
<td>2</td>
<td>Rohit Girdhar</td>
<td>DeepMind</td>
<td>21.03</td>
</tr>
<tr>
<td>3</td>
<td>Ting Yao</td>
<td>YH Technologies Co., Ltd.</td>
<td>19.60</td>
</tr>
<tr>
<td>4</td>
<td>George Lee</td>
<td>Fudan</td>
<td>17.16</td>
</tr>
<tr>
<td>5</td>
<td>Xiyang Dai</td>
<td>UMD</td>
<td>16.70</td>
</tr>
<tr>
<td>6</td>
<td>Peppa Pig</td>
<td>For ECCV</td>
<td>13.56</td>
</tr>
<tr>
<td>7</td>
<td>Ho Ran</td>
<td>Ran Ho</td>
<td>13.46</td>
</tr>
<tr>
<td>8</td>
<td>Ke Yun Yun</td>
<td>Yun Ke</td>
<td>13.05</td>
</tr>
<tr>
<td>9</td>
<td>Kevin Lin</td>
<td>University of Washington</td>
<td>12.25</td>
</tr>
<tr>
<td>10</td>
<td>Oytun Ulutan</td>
<td>UCSB</td>
<td>11.36</td>
</tr>
<tr>
<td>11</td>
<td>Gurkirt Singh</td>
<td>Oxford Brookes University</td>
<td>9.42</td>
</tr>
<tr>
<td>12</td>
<td>cliff wang</td>
<td>LW</td>
<td>7.81</td>
</tr>
<tr>
<td>13</td>
<td>x G</td>
<td>BLWC</td>
<td>7.81</td>
</tr>
<tr>
<td>14</td>
<td>Bin Wang</td>
<td>Little Wheel Co.</td>
<td>0.66</td>
</tr>
</tbody>
</table>
For context: Winning team architecture
Newest model:

Video Action Transformer Network

Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman

(Submitted on 6 Dec 2018 (v1), last revised 17 May 2019 (this version, v2))

Input clip (RGB frames)

I3D (3D conv trunk)

Initial video representation

RPN

Potential actor locations

RolPool

Preprocessing

Initial actor representation

Is this actor representation sufficient to recognize actions?
ActionTransformer block: person-specific self attention

Repurposing the Transformer (NIPS’17) for Spatiotemporal Action Detection

Initial video representation → RolPool → Preprocessing

Initial actor representation → Softmax Attention → Weighted Sum → More transformers

Updated actor representation

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>11.3</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>15.6</td>
</tr>
<tr>
<td>Better baseline</td>
<td>21.0</td>
</tr>
<tr>
<td>Action transformer</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Vaswani et al. *Attention is all you need.* NIPS’17
Similar ideas also explored in Sun et al. *Actor Centric Relation Networks.* ECCV’18
Conclusions

- Action recognition dataset where models trained on it may have directly practical applications (unlike Kinetics)
- Performance still rather low (but good improvements this year: check ActivityNet’s workshop tomorrow)
- Lots of research opportunities
Thank you!