Transferring from Kinetics

João Carreira

Tutorial on Action Classification and Video Modelling
CVPR 2019
16th of June

DeepMind

Brief Recent History of Image Understanding

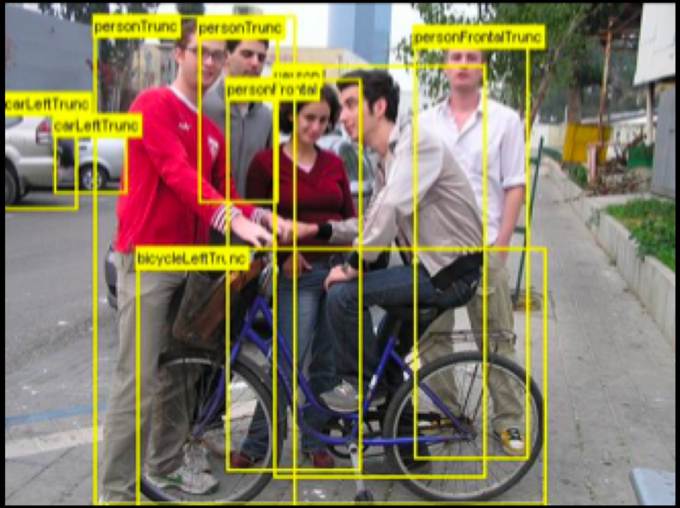
Caltech 101 (10k images):
- classification

2001-2012

PASCAL VOC (10k images):
- object localization

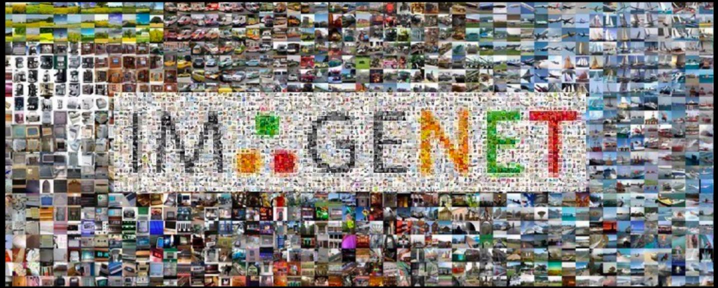
Simple features

Problems studied in isolation



Brief Recent History of Image Understanding

ImageNet (1M images):
- classification



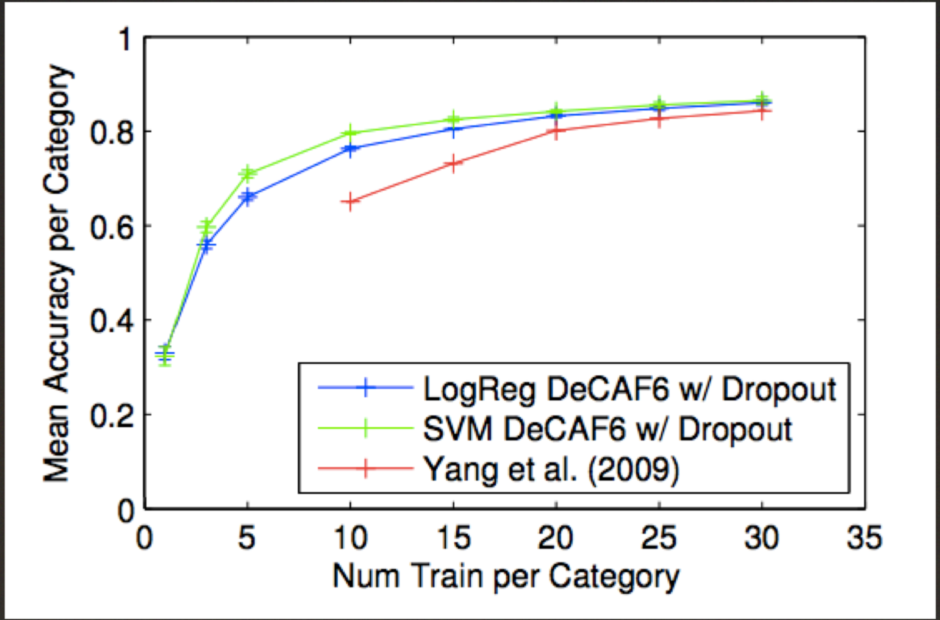
2012-now

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

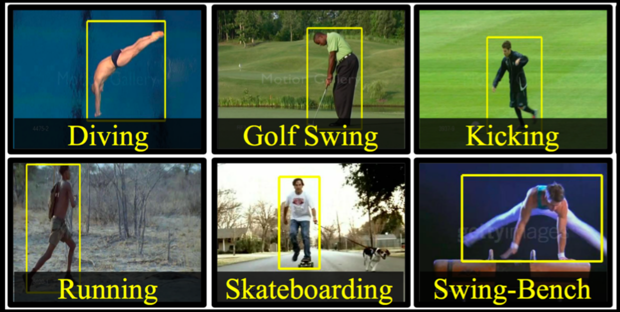
UCF 101 (10k videos) / HMDB-51 (5k videos):
- classification

2012-2016

ActivityNet, Thumos, UCF101-Det:
- Action localization

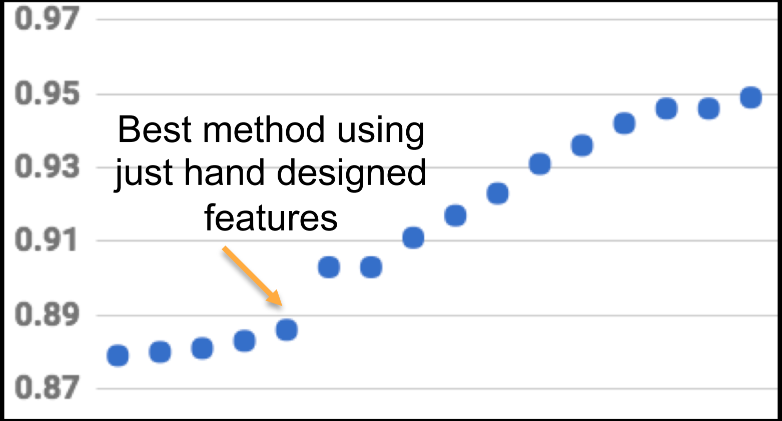
Problems studied in isolation

Transfer from ImageNet

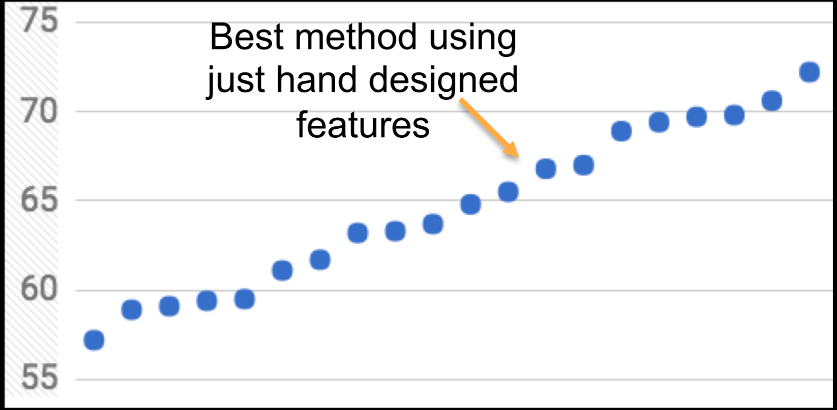


Transferring from ImageNet to Video

UCF-101



HMDB-51



Compilation of results from actionrecognition.net

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

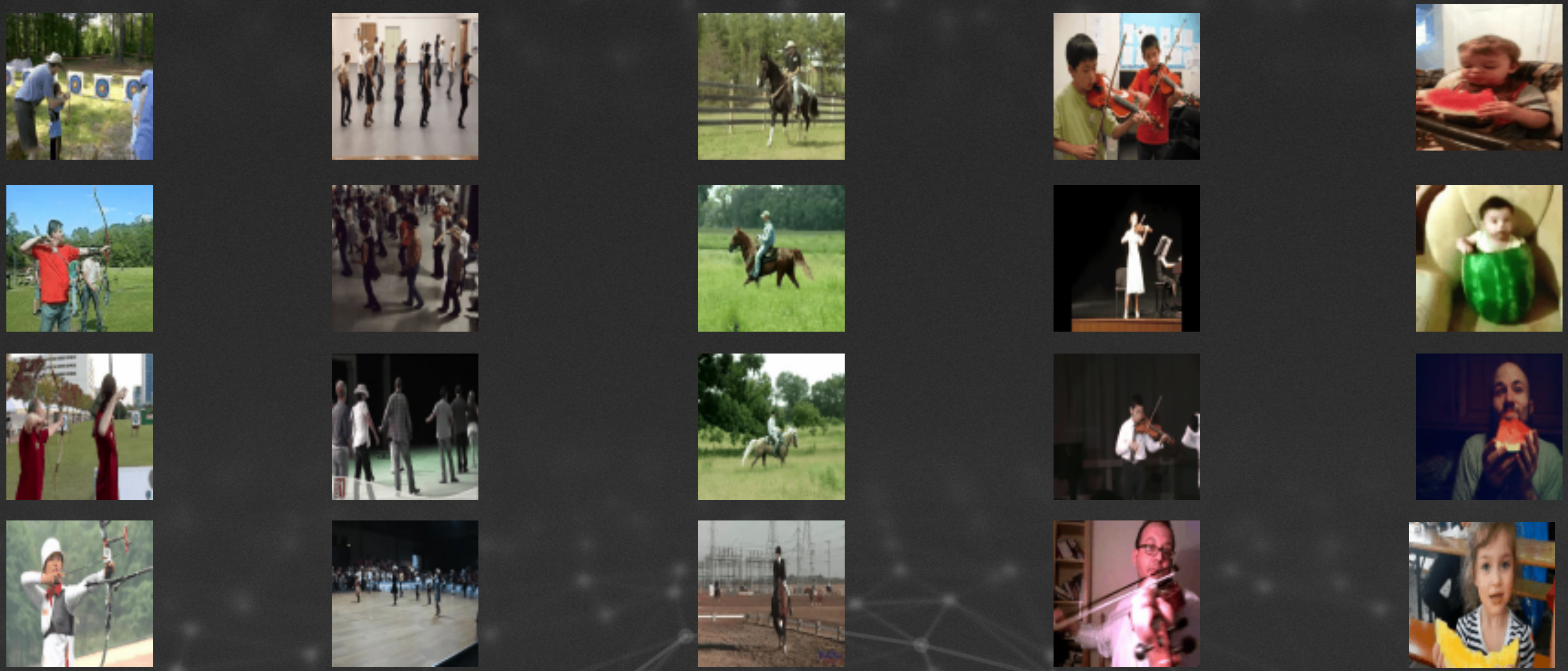
2017

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)

~~ImageNet~~ **Kinetics**

~~Object classification~~

Human action classification (10s clips)

~~ImageNet~~ **Kinetics**

~~1000 object classes x 1000 images~~

400 human action classes x >400 videos

(300k total, ~all from unique videos)

~~ImageNet~~ **Kinetics**

~~Images from google searches~~

Videos from youtube searches

Previous human action classification datasets too tiny to properly research new video representations

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500
ActivityNet-200 [3]	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245

Dataset Collection

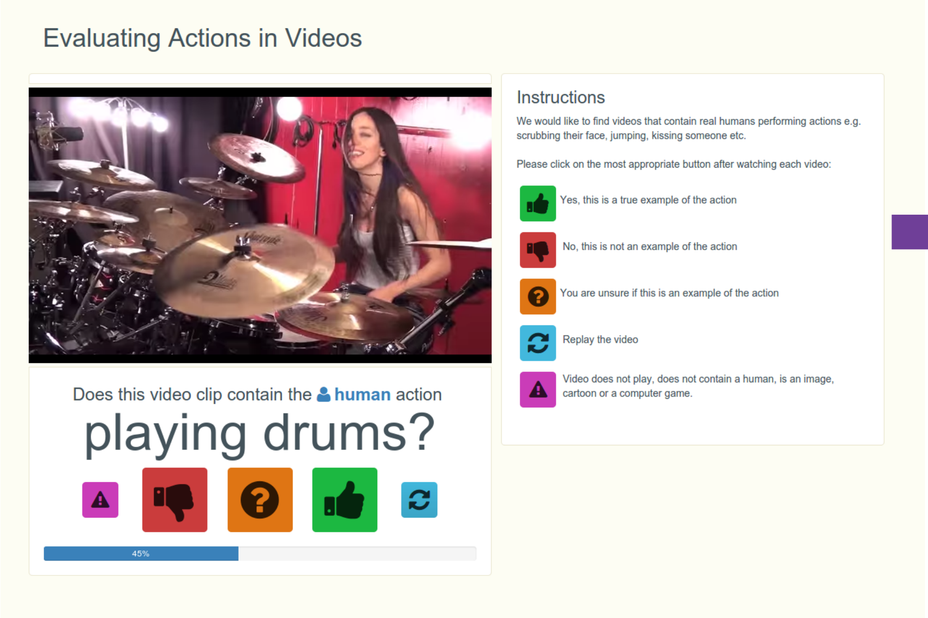
0 abseiling
1 laughing
2 swimming
3 shearing sheep
4 motorcycling
5 celebrating
6 spray painting
7 playing tennis
8 driving tractor
9 washing dishes
10 skateboarding
11 waxing legs

Title matching

How to make healthy eating unbelievably easy | Luke
TEDx Talks

Image Classifiers

Human verification using **Mechanical Turk**



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

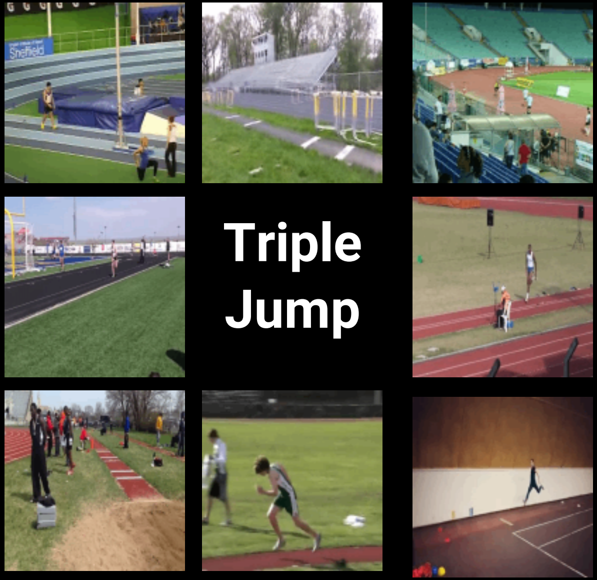
Person Actions (Singular)


Pumping Fist


Shaking Head

Action list

Person Actions (Singular)



Action list

Person-person actions

Action list

Person-object actions

Action list

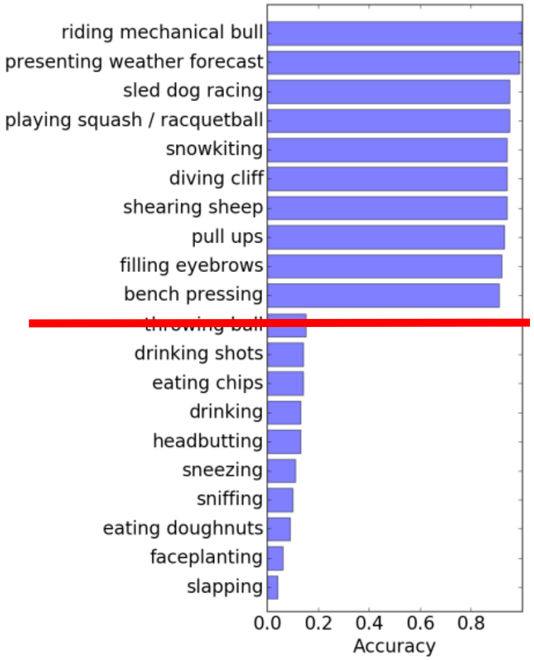
Person-object actions

Folding Clothes

Folding Napkin

Action list

Person-object actions



DeepMind Shows AI Has Trouble Seeing Homer Simpson's Actions

By Jeremy Hsu
Posted 8 Jun 2017 | 14:00 GMT

Image: FOX/Getty Images

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

Technology | Innovation

Homer Simpson defeats Google's all-powerful DeepMind artificial intelligence

Super computer not smart enough to visually recognise many of Homer's signature actions.

By Mary-Ann Russon
June 12, 2017 11:29 BST

Google DeepMind computer scientists say artificial intelligence is still struggling to comprehend common Homer Simpson actions like drinking beer and eating donuts (20th Century Fox)

D'oh! You'd never believe it, but in a new research paper, computer scientists at Google DeepMind have admitted that its artificial intelligence technology still struggles to identify many common human behaviours that Homer Simpson exhibits – whether it's eating doughnuts or crisps, falling on his face, yawning or drinking beer.

Kinetics has kept growing

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500
ActivityNet-200 [3]	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245
Kinetics-600	**2018**	**600**	**min 450**	**500,000**	**500,000**
Kinetics-700	**2019**	**700**	**min 450**	**650,000**	**650,000**

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

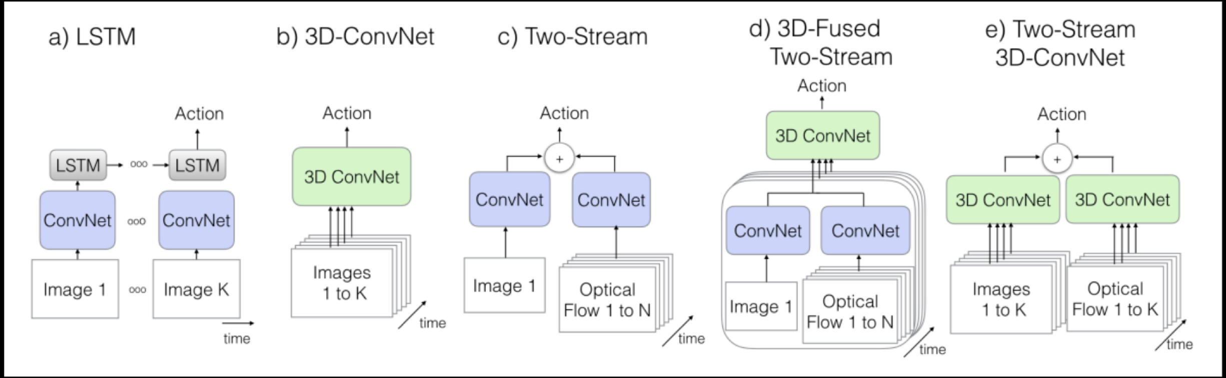
ActivityNet, Charades, AVA
- Action localization

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



Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet (C3D)	79M	16 rgb	0.64s	240 rgb	9.6s
Two-Stream	12M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
3D-Fused	39M	5 rgb, 50 flow	2s	25 rgb, 250 flow	10s
Two-Stream I3D	25M	64 rgb, 64 flow	2.56s	250 rgb, 250 flow	10s

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

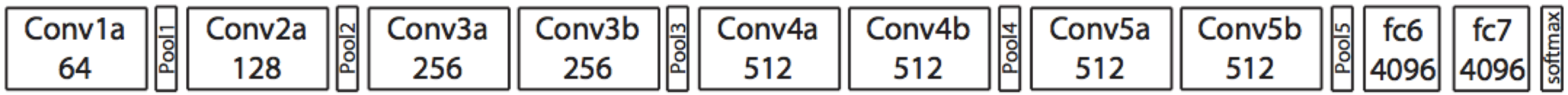


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling 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

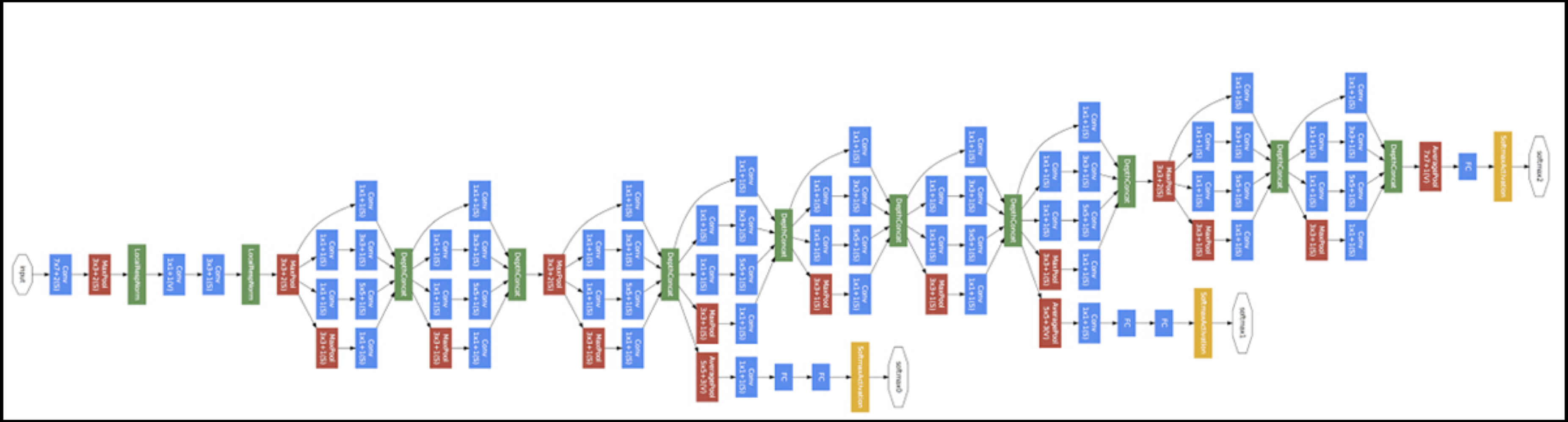
Pure video model, that learns a hierarchical representation directly over video

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

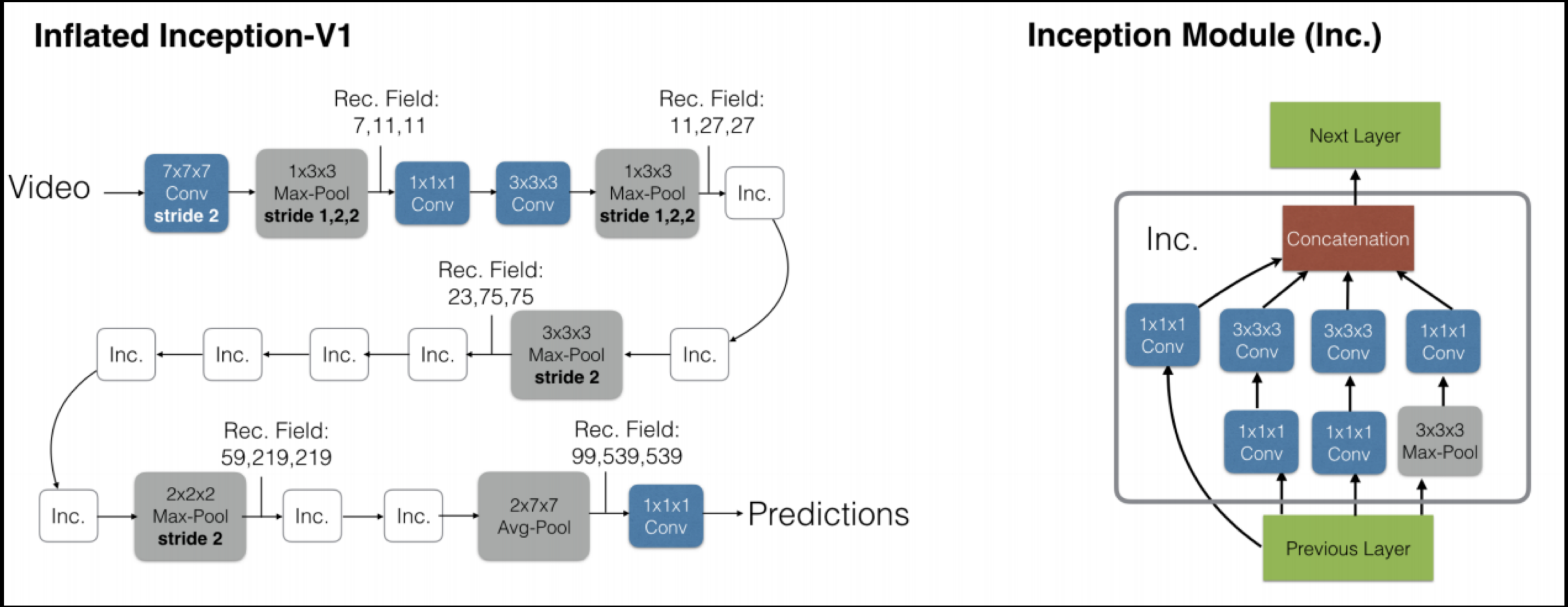
C3D (1 net) + linear SVM	82.3
C3D (3 nets) + linear SVM	**85.2**
Two-stream networks [36]	88.0

Google's Inception-V1 ImageNet classifier

Going deeper with convolutions, Szegedy et al, CVPR 2015

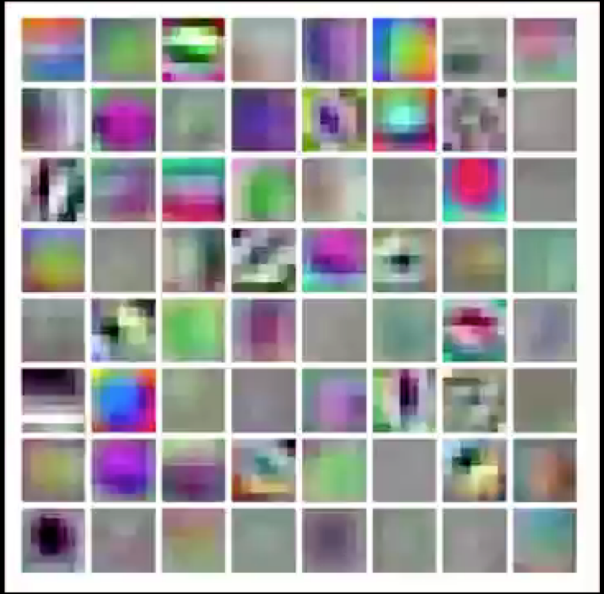


Inflated 3D Inception (I3D)

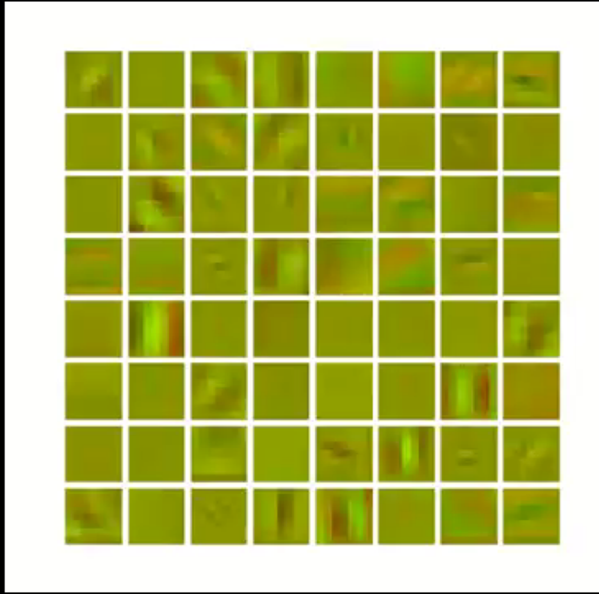


I3D Conv1 filters, trained in Kinetics

RGB



Flow

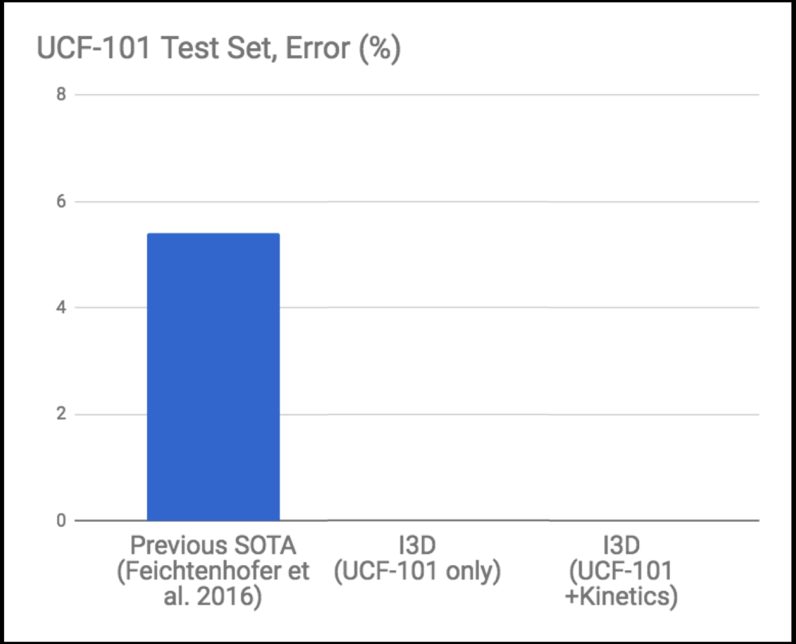


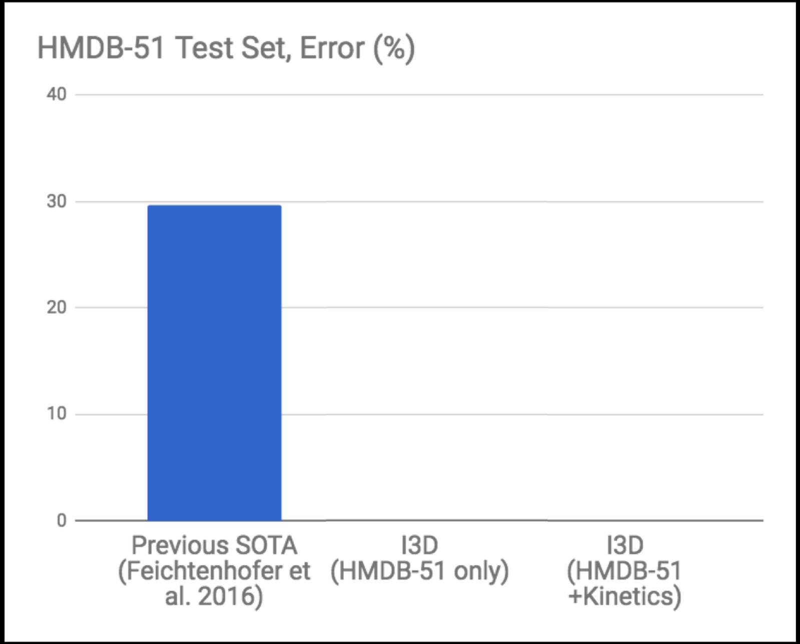
Transfer results with miniKinetics pre-training (80k videos)

Architecture	UCF-101				HMDB-51			
	Original	Fixed	Full-FT	Δ	Original	Fixed	Full-FT	Δ
(a) LSTM	81.0	81.6	82.1	-6%	36.0	46.6	46.4	-16.7%
(b) 3D-ConvNet (C3D)	49.2	76.0	79.9	**-60.5%**	24.3	47.5	49.4	**-33.1%**
(c) Two-Stream	91.2	90.3	91.5	-3.4%	58.3	64.0	58.7	-13.7%
(d) 3D-Fused	89.3	88.5	90.1	-7.5%	56.8	59.0	61.4	-10.6%
(e) Two-Stream I3D	**93.4**	**95.7**	**96.5**	-47.0%	**66.4**	**74.3**	**75.9**	-28.3%

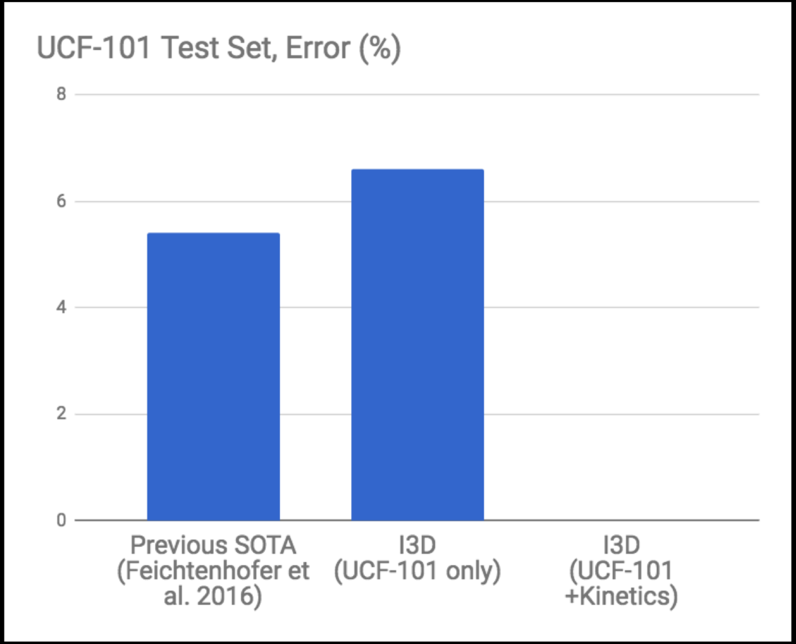
Table 3. Performance on the UCF-101 and HMDB-51 test sets (splits 1 of both) for architectures pre-trained on miniKinetics. All except 3D-ConvNet are based on Inception-v1 and start off pre-trained on ImageNet. Original: train on UCF-101 / HMDB-51; Fixed: features from miniKinetics, with the last layer 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.

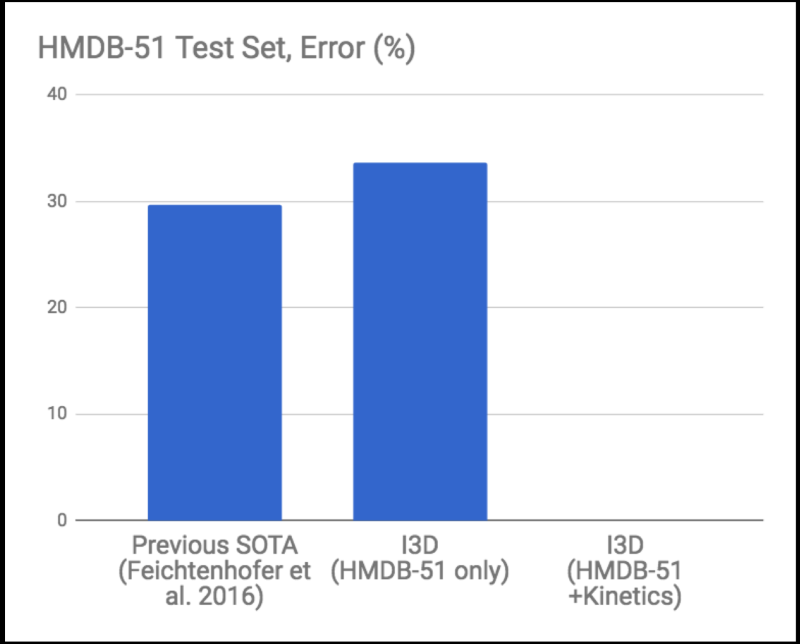
Comparison with state-of-the-art



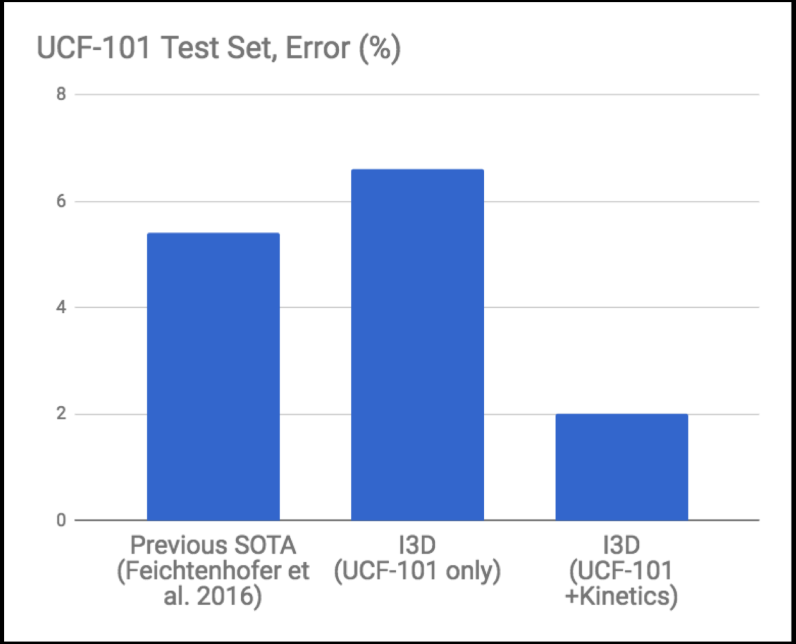


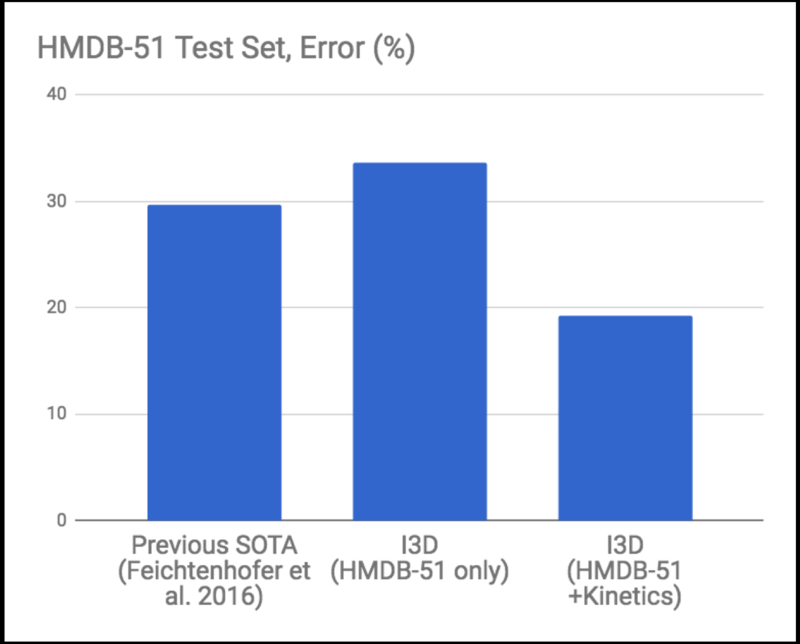
Comparison with state-of-the-art



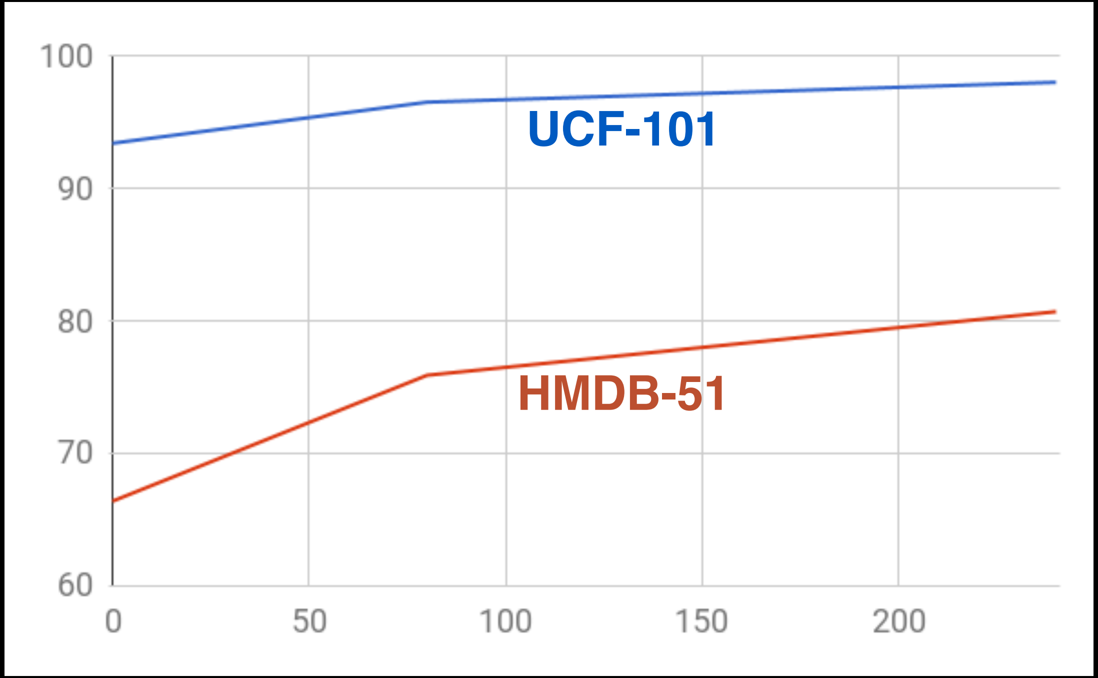


Comparison with state-of-the-art



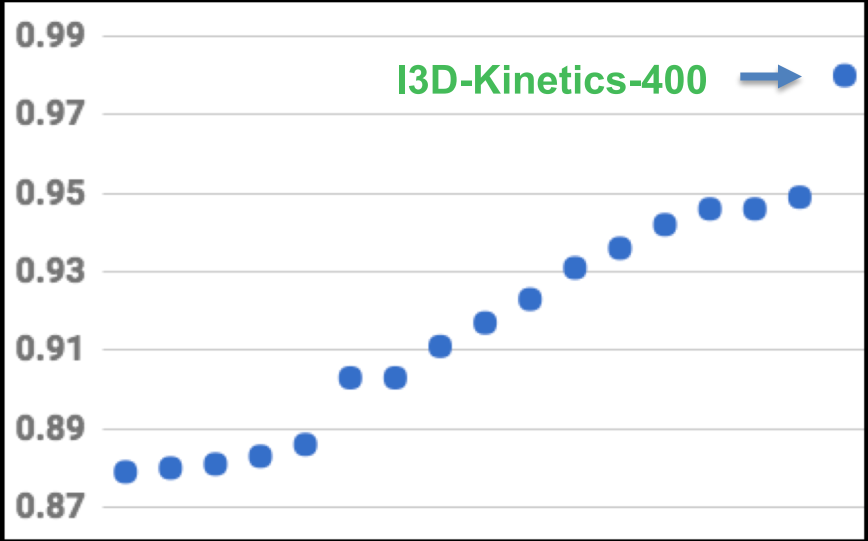


Performance as function of # Kinetics examples

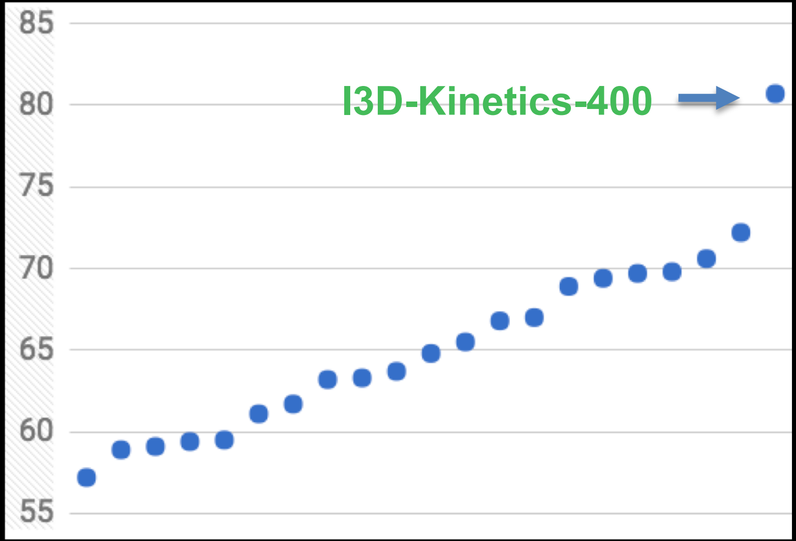


I3D-Kinetics-400 transfer performance (two stream, flow+rgb)

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

Action Recognition Results

Rank	Team	Accuracy (mAP)	Modeling Approach
1	TeamKinetics	0.3441	I3D ConvNet with dense per-frame outputs
2	DR/OBU	0.2974	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.
3	UMICH-VL	0.2811	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.

Charades challenge winning entry at CVPR 2017

Temporal Segmentation Results

Rank	Team	Accuracy (mAP)	Modeling Approach
1	TeamKinetics	0.2072	I3D ConvNet with dense per-frame outputs
2	UMICH-VL	0.1803	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.
3	DR/OBU	0.1796	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.

Charades dataset

Publications

1. *The Kinetics Human Action Video Dataset*. Kay, Carreira, Simonyan, Zhang, Hillier, Vijayanarasimhan, Viola, Green, Back, Natsev, Suleyman and Zisserman, arXiv 2017.
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

Google AI

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

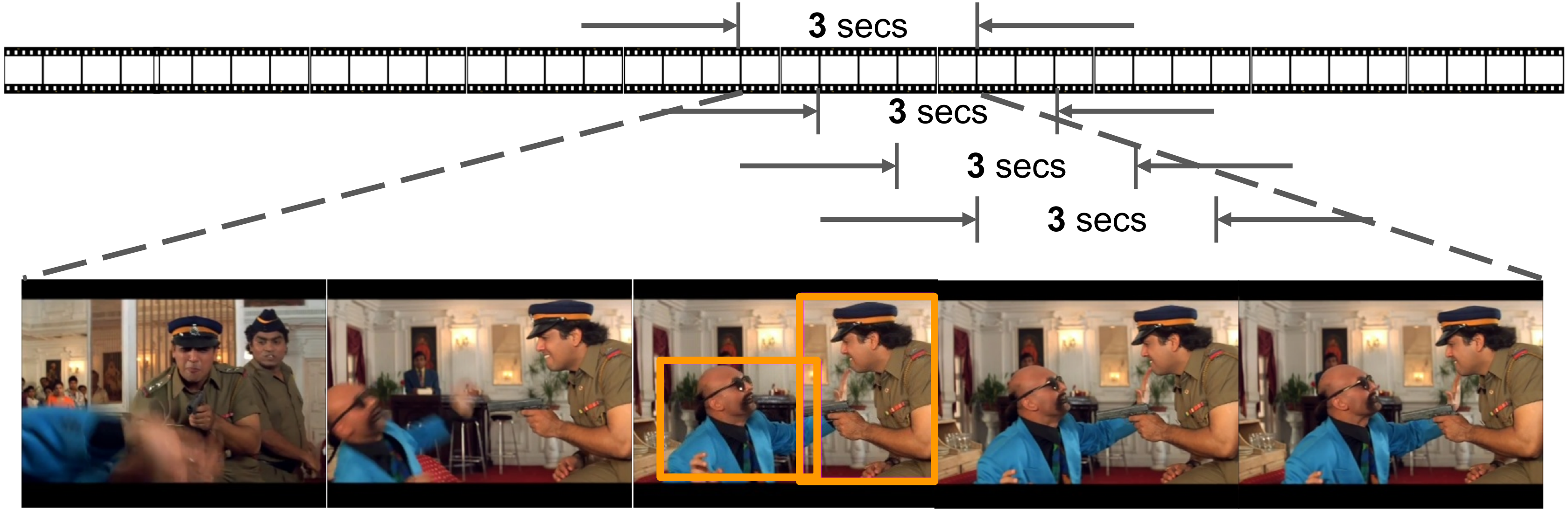
Pose (14)
run/jog
walk
jump
stand
sit
lie/sleep
bend/bow
crawl
swim
dance
get up
fall down
crouch/kneel
martial art

Person-Person (17)
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-Object (49)

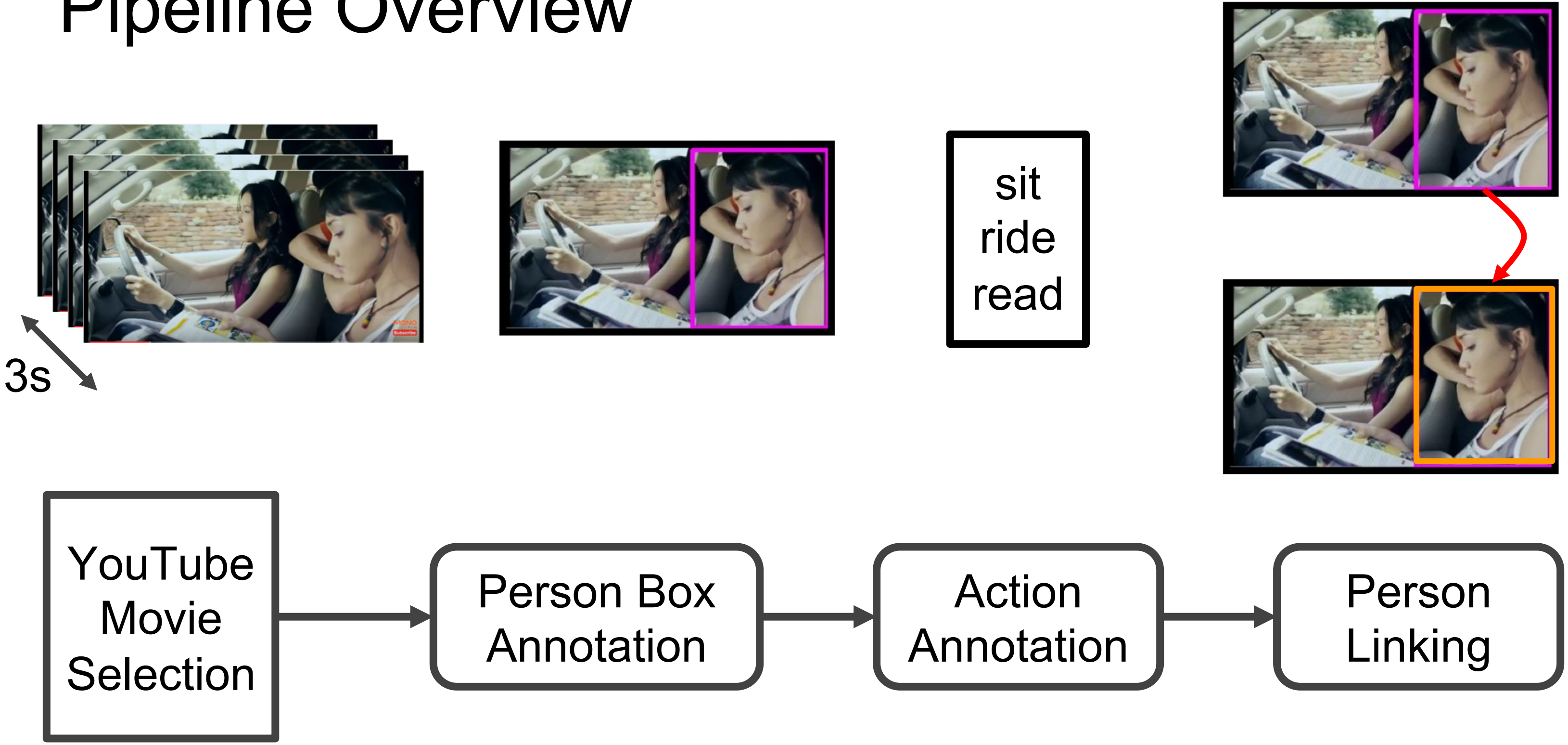
lift/pick up	smoke	work on a computer	open
put down	sail boat	answer phone	close
carry	row boat	climb (e.g., mountain)	enter
hold	fishing	play board game	exit
throw	touch	play with pets	
catch	cook	drive (e.g., a car)	
eat	kick	push (an object)	
drink	paint	pull (an object)	
cut	dig	point to (an object)	
hit	shovel	play musical instrument	
stir	chop	text on/look at a cellphone	
press	shoot	turn (e.g., screwdriver)	
extract	take a photo	dress / put on clothing	
read	brush teeth	ride (e.g., bike, car, horse)	
write	clink glass	watch (e.g., TV)	

Atomicity from 3-sec segment sampled at 1Hz



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

Pipeline Overview

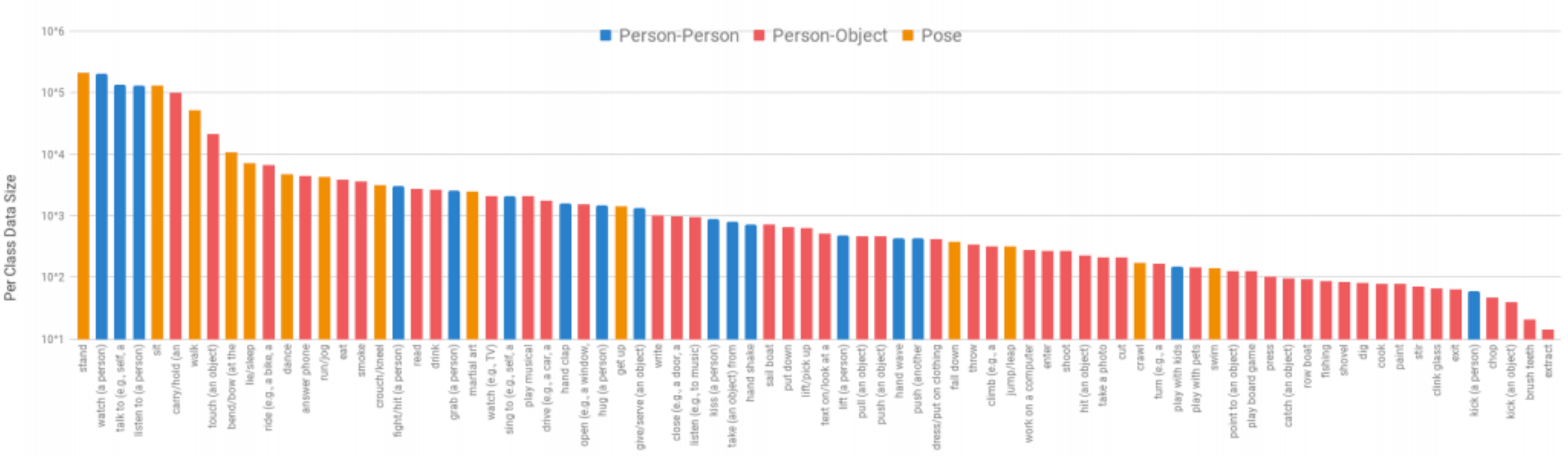


Dataset Statistics

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

First Action	Second Action	NPMI
Watch (TV/monitor)	Work on a computer	0.64
Open (window/door)	Close (door/box)	0.59
Text on/Look at a cell phone	Answer phone	0.53
Listen to (a person)	Talk to (a person)	0.47
Fall down	Lie/Sleep	0.46
Talk to (a person)	Listen to (a person)	0.43
Stand	Sit	0.40
Walk	Stand	0.40

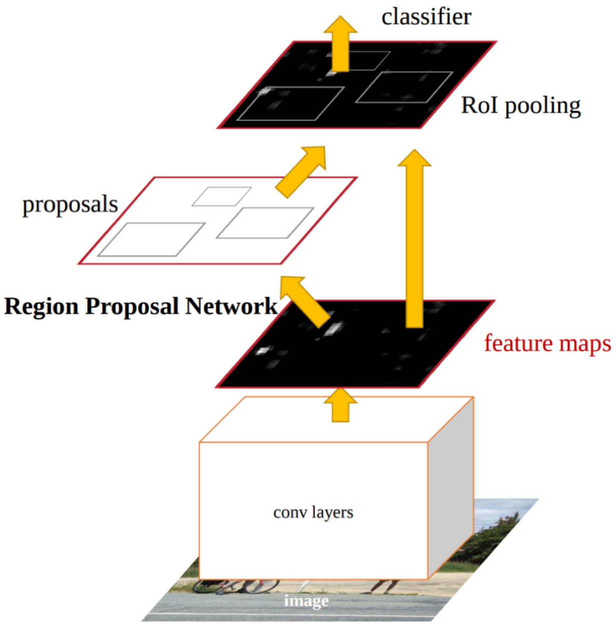
Action Co-occurrence among Persons

Person 1 Action	Person 2 Action	NPMI
Ride (bike/car/horse)	Drive (car/truck)	0.60
Play musical instrument	Listen to (music)	0.57
Take (object)	Give/Serve (object)	0.51
Talk to (a person)	Listen to (a person)	0.46
Stand	Sit	0.31
Play musical instrument	Dance	0.23
Watch (a person)	Write	0.15
Walk	Run/Jog	0.15

Baseline Performance

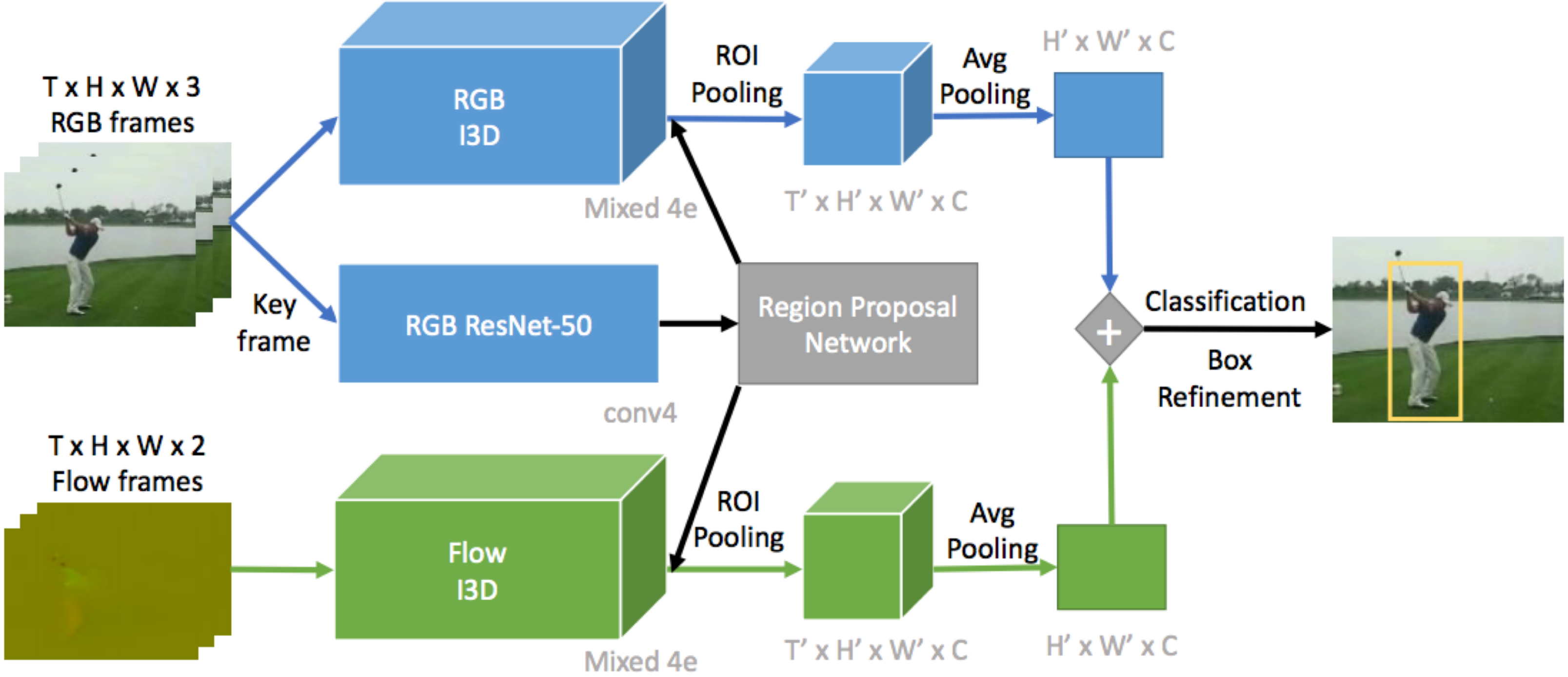
Original Baseline 1

Faster R-CNN with ResNet-101 from ImageNet



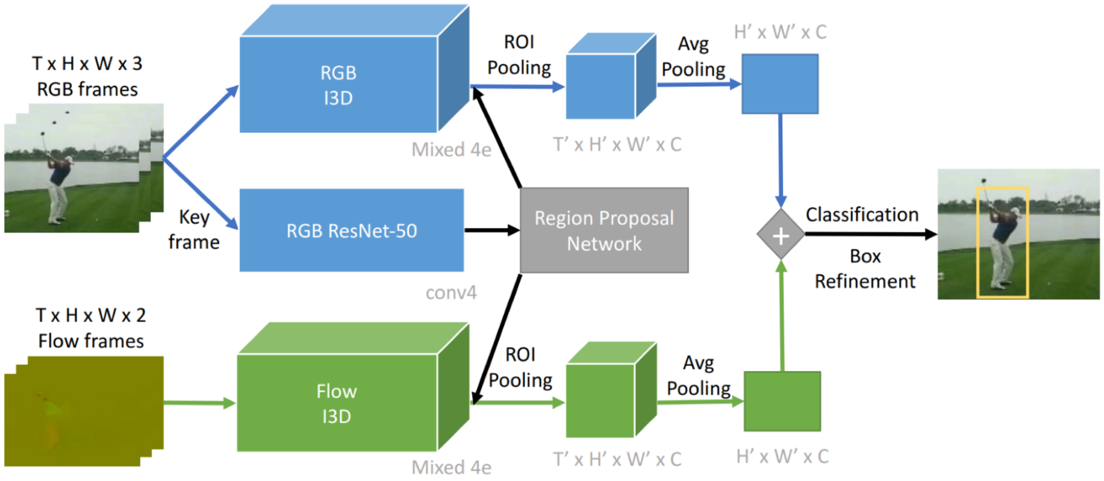
Method	mAP
Baseline 1	11.3

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)



Method	mAP
Baseline 1	11.3
Baseline 2	15.6

AVA challenge 2018:

arXiv.org > cs > arXiv:1807.10066

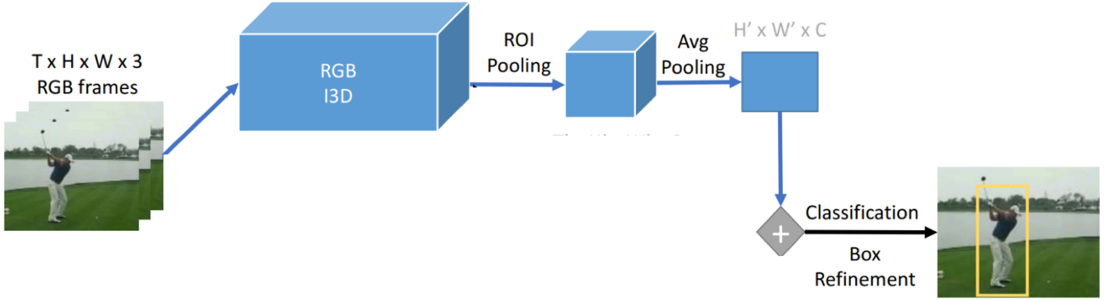
Computer Science > Computer Vision and Pattern Recognition

A Better Baseline for AVA

Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman

(Submitted on 26 Jul 2018)

RGB I3D (in Faster R-CNN framework)



Method	mAP
Baseline 1	11.3
Baseline 2	15.6
Ours	21.0

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

AVA challenge 2018

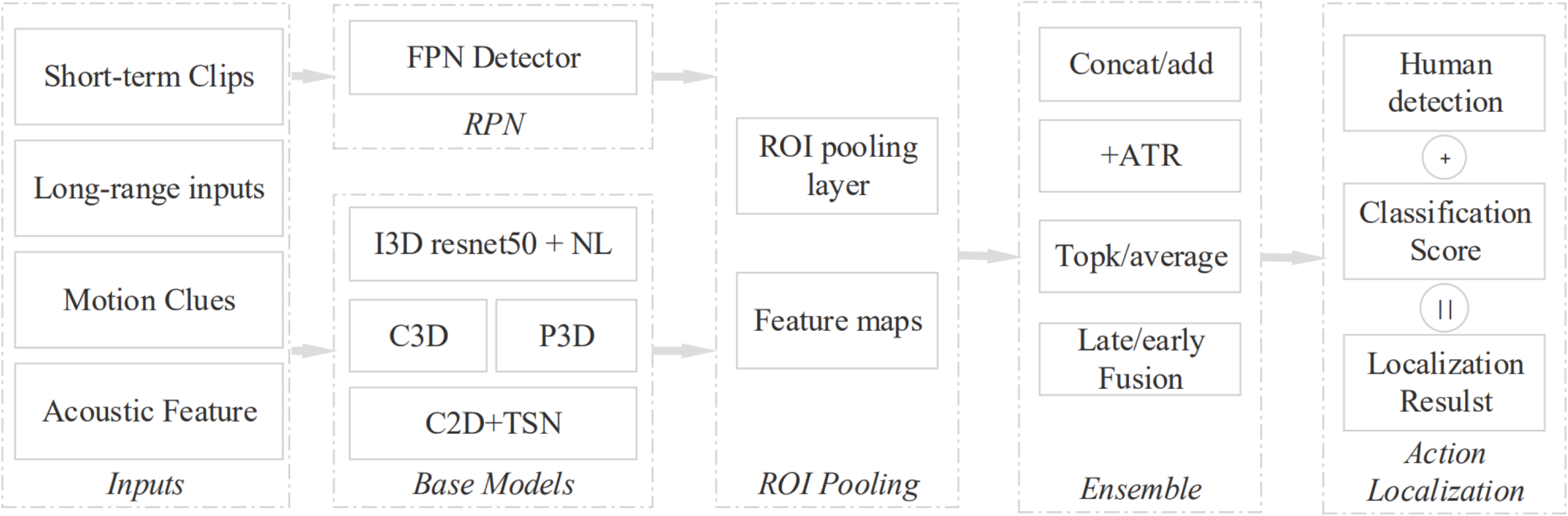
Jianwen Jiang[1], Yu Cao[2], Lin Song[3], Shiwei Zhang[4] Yunkai Li[5], Ziyao Xu[5], Qian Wu[6],
Chuang Gan[1*], Chi Zhang[5*], Gang Yu[5*]
[1]Tsinghua University, jjw17@mails.tsinghua.edu.cn, ganchuang1990@gmail.com
[2]Beihang University, cqcy1208@buaa.edu.cn
[3]Xian Jiaotong University, stevengrove@xtu.xjtu.edu.cn
[4]Huazhong University of Science and Technology, swzhang@hust.edu.cn
[5]Megvii Inc. (Face++), {liyunkai, xuziyao, zhangchi, yugang}@megvii.com
[6]Zhejiang University, wq1601@zju.edu.cn

Task #1 - Computer Vision

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

baseline performance 15.6

For context: Winning team architecture



Newest model:

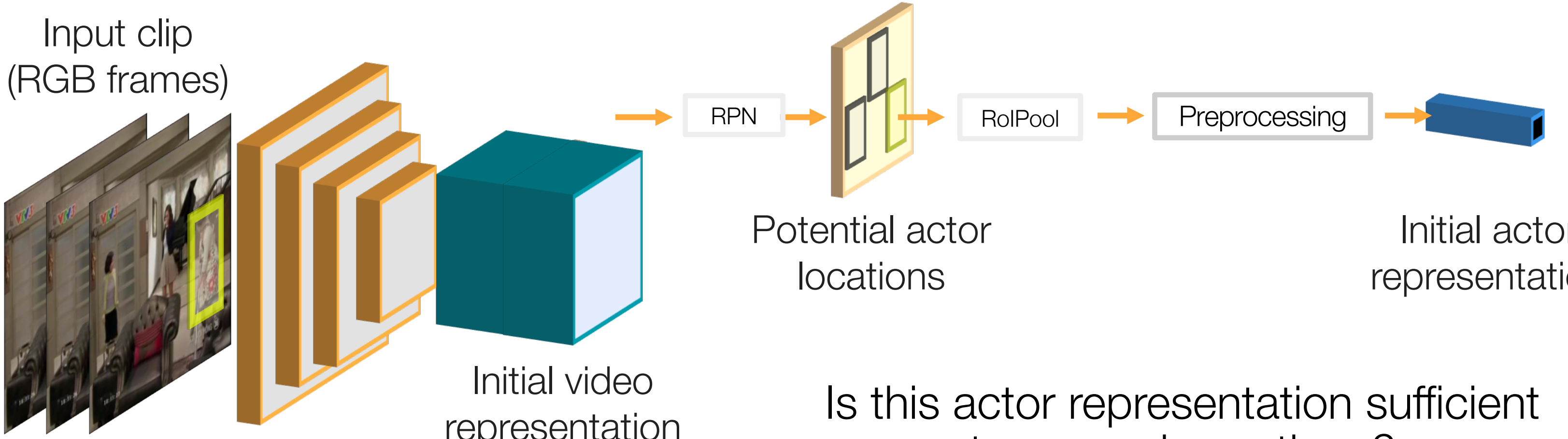

arXiv.org > cs > arXiv:1812.02707

Computer Science > Computer Vision and Pattern Recognitio

Video Action Transformer Network

Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman

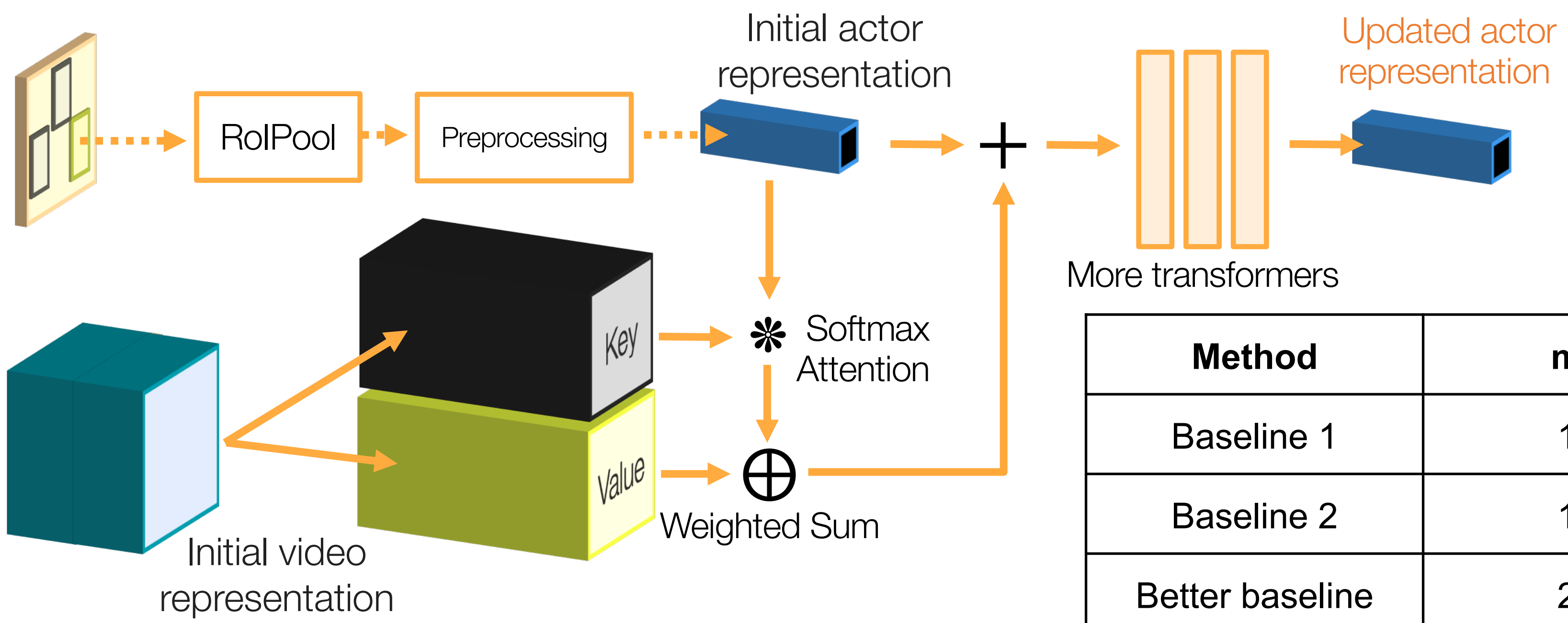
(Submitted on 6 Dec 2018 (v1), last revised 17 May 2019 (this version, v2))



Is this actor representation sufficient to recognize actions?

ActionTransformer block: person-specific self attention

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



Method	mAP
Baseline 1	11.3
Baseline 2	15.6
Better baseline	21.0
Action transformer	24.9

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!