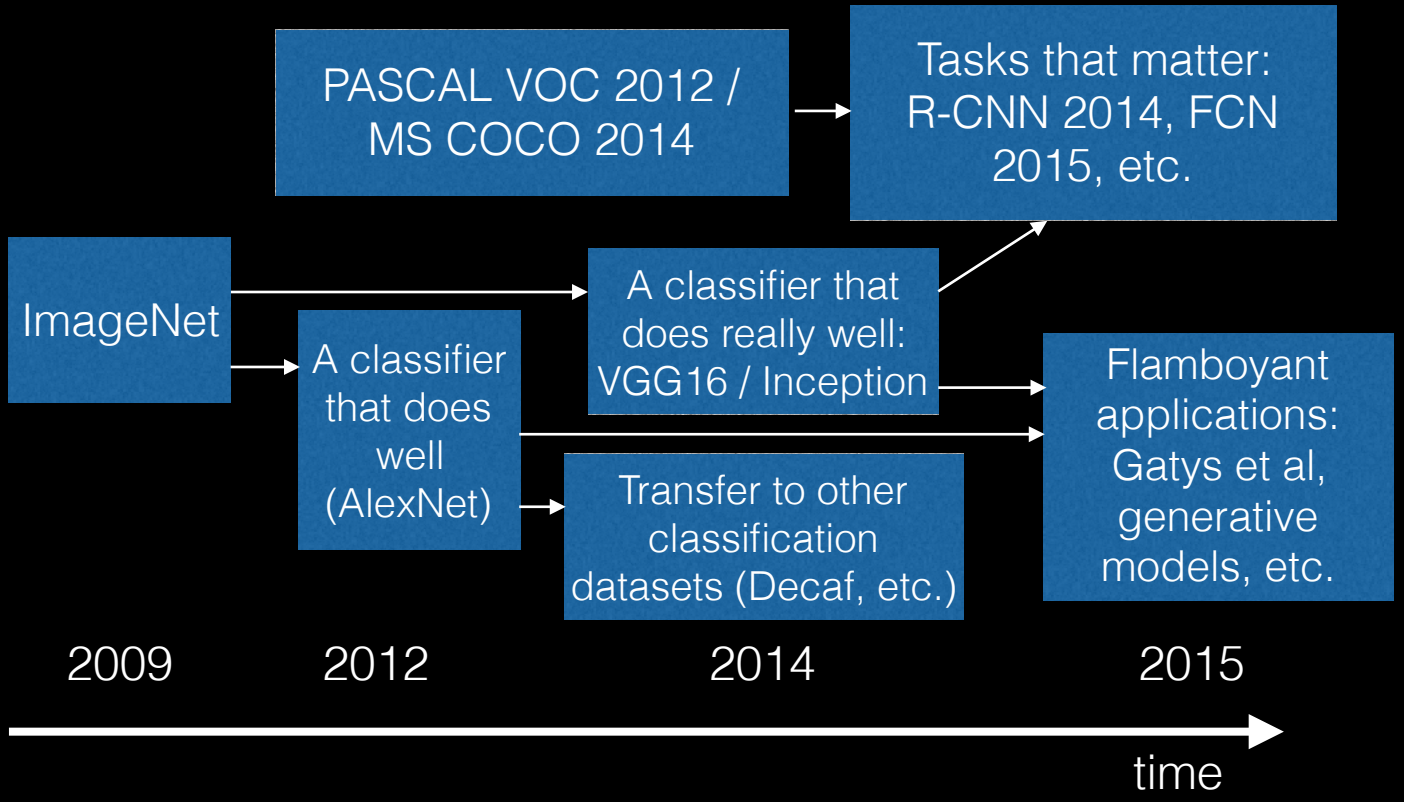
Video Understanding and the Kinetics Dataset

João Carreira

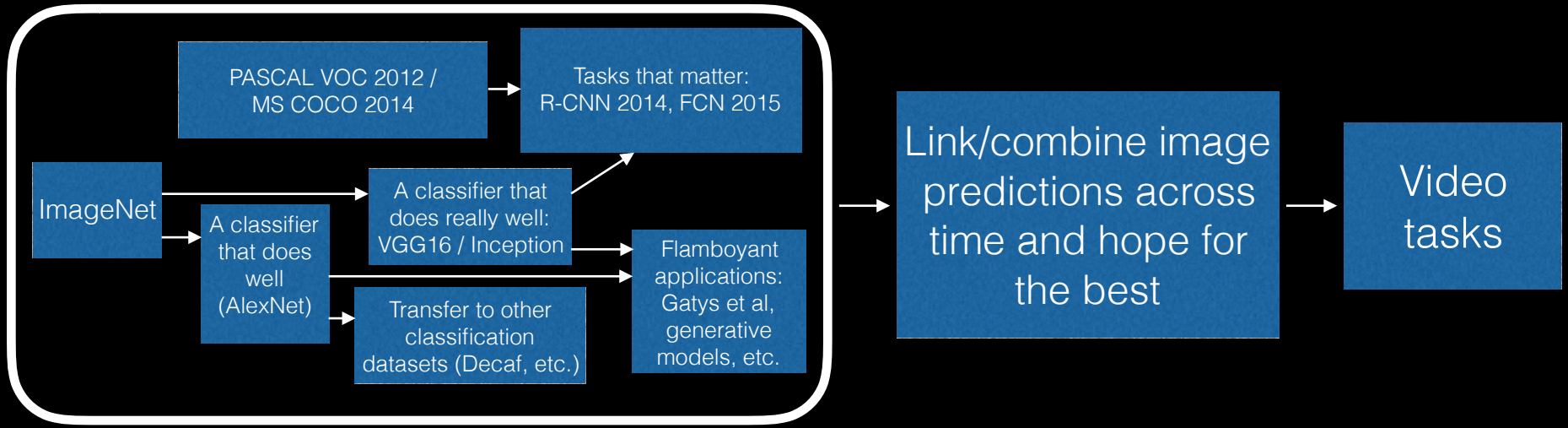
Joint work with Andrew Zisserman, Brian Zhang, Will Kay, Karen Simonyan, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman

26th of July, 2017
ActivityNet: Large-scale Activity Recognition Challenge

Image understanding

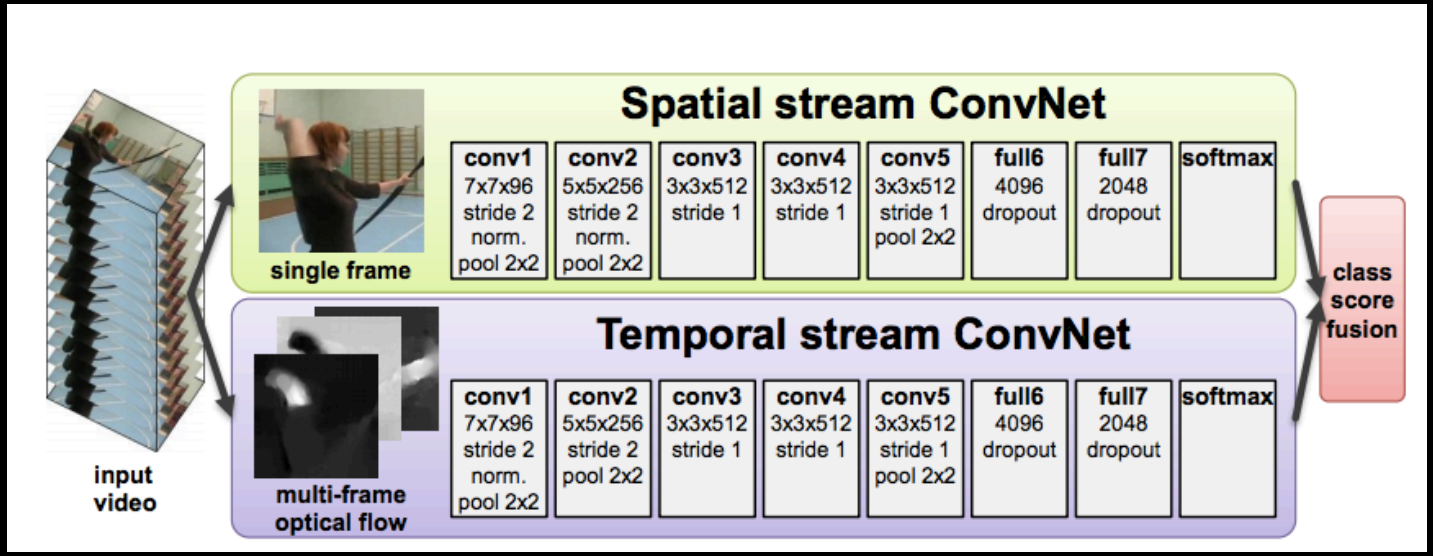


Video understanding: dominant strategy so far



Modern video neural network model: Two-stream Networks

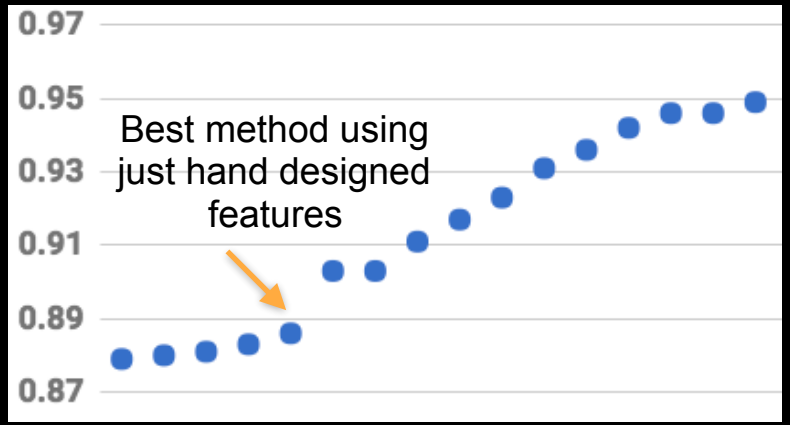
Largely based on averaging ImageNet image classifier predictions across multiple frames



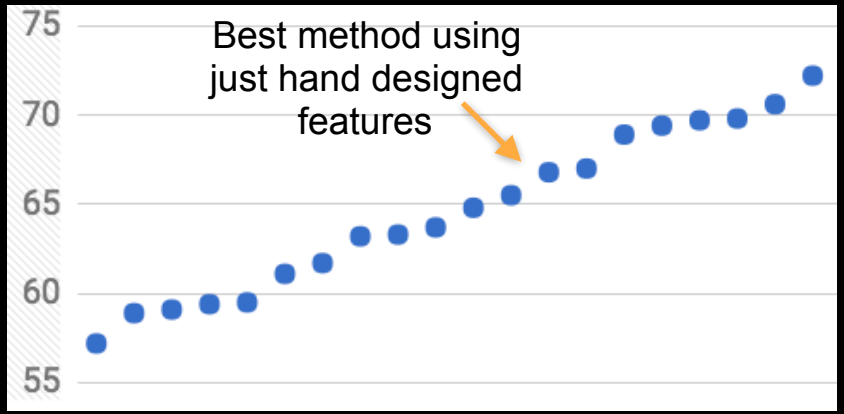
Simonyan and Zisserman, 2014

Evolution of the state-of-the-art

UCF-101



HMDB-51



Compilation of results from actionrecognition.net

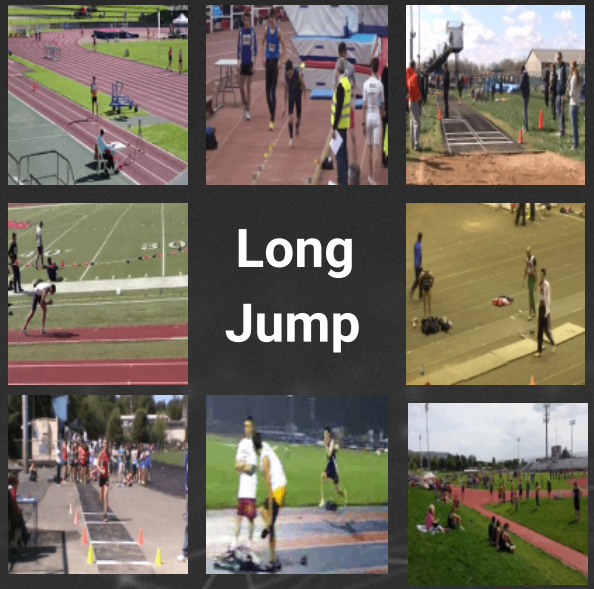
Problems - 1 of 2

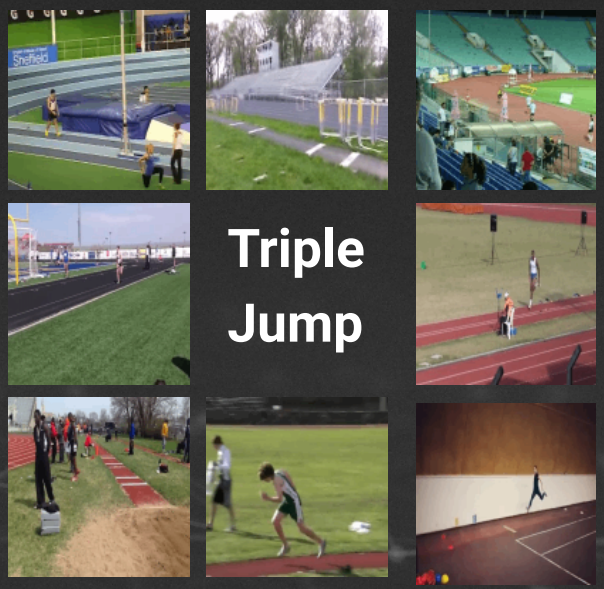
How to learn hierarchical motion features ?

Gunnar Johannson, video from 1971

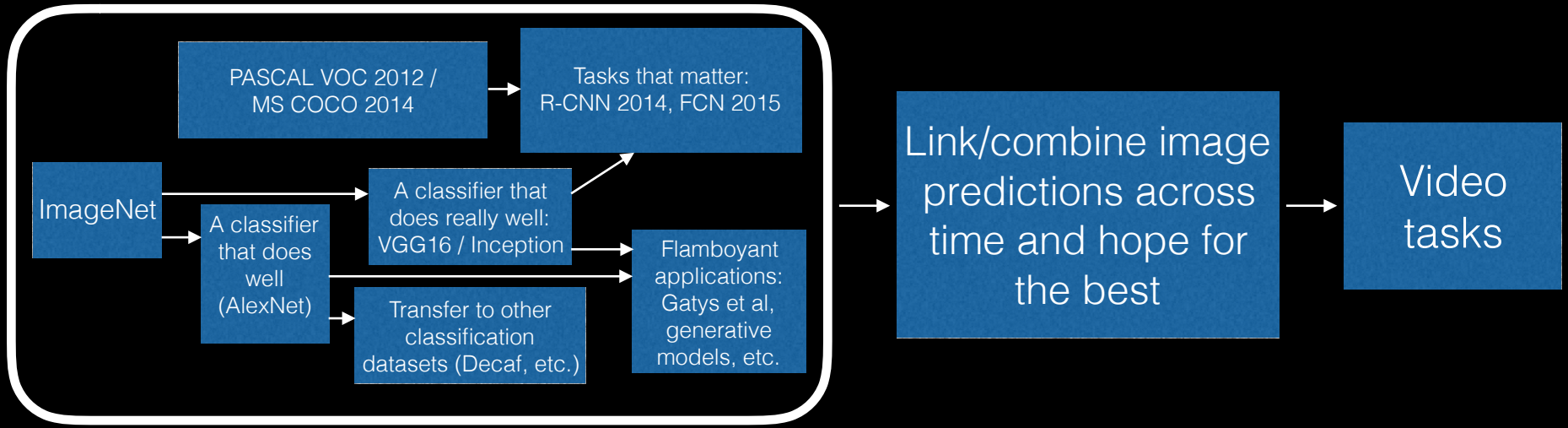
Problems - 2 of 2

Even if one frame is informative enough, how can we find it and suppress noisier predictions from the other frames ?

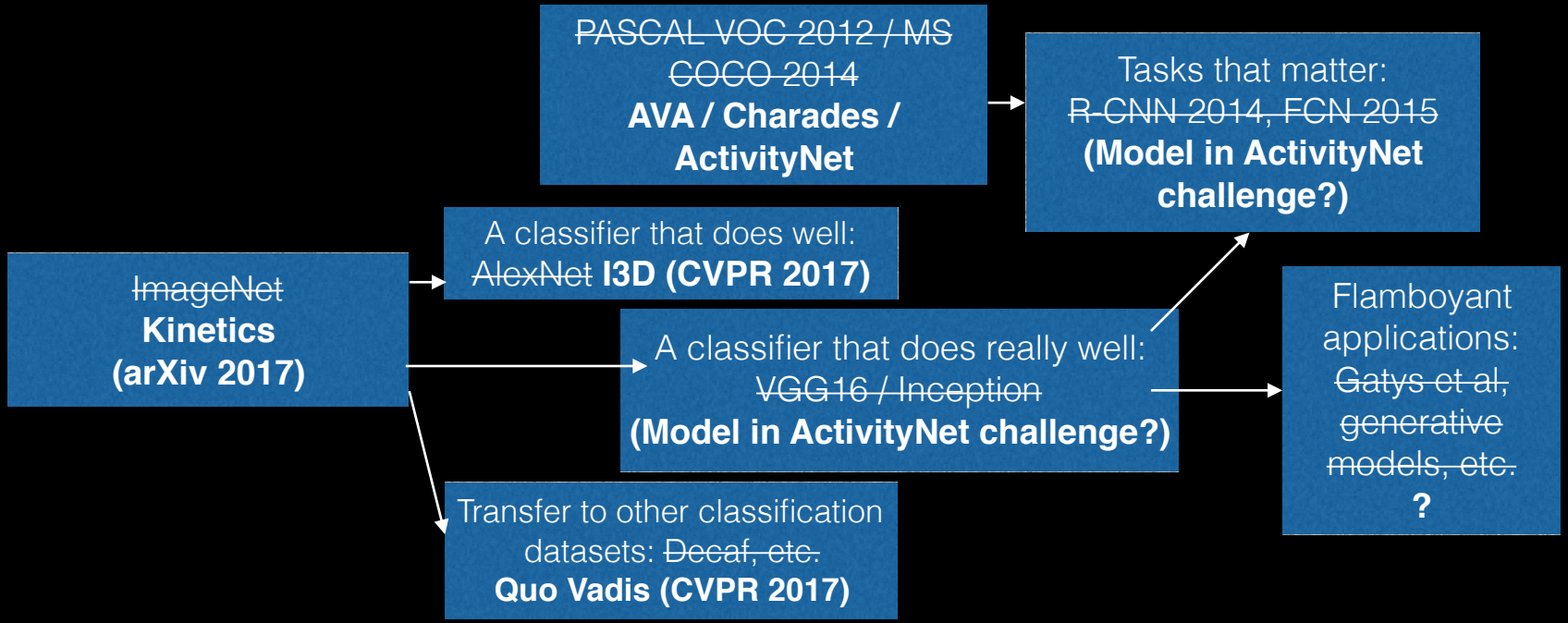




Video understanding: dominant strategy so far

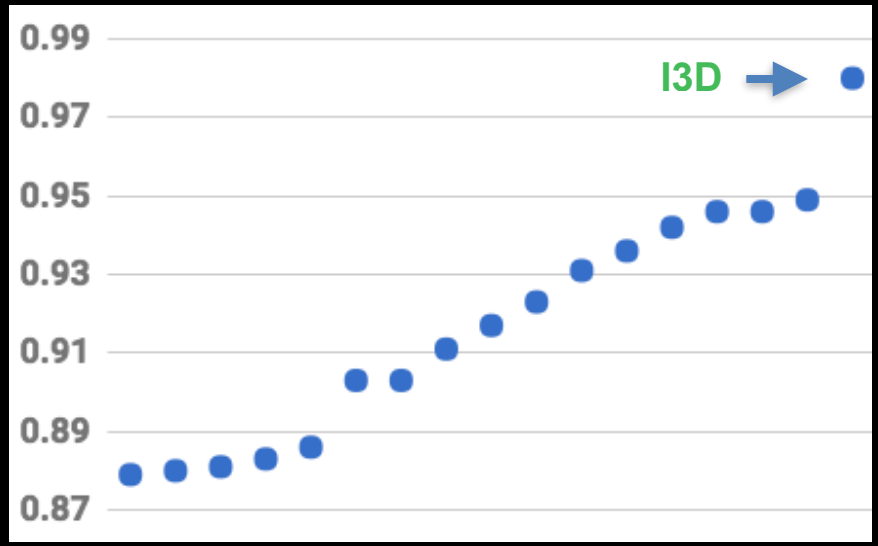


Alternative plan for video understanding

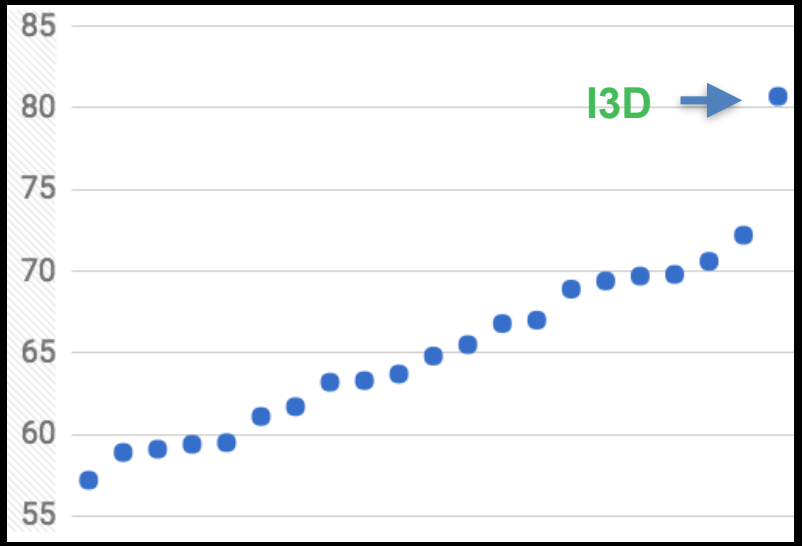


I3D-Kinetics transfer performance (two stream, flow+rgb)

UCF-101

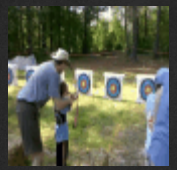
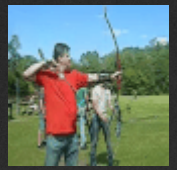
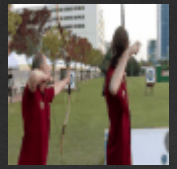


HMDB-51



Kinetics pre-training, comparison with state-of-the-art (compilation of results from actionrecognition.net)

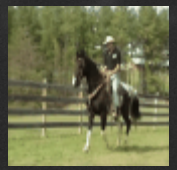
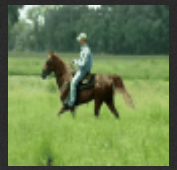
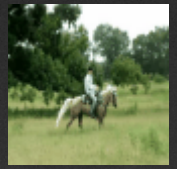
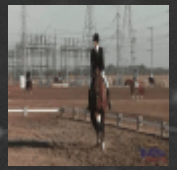
1. The Kinetics dataset


archery

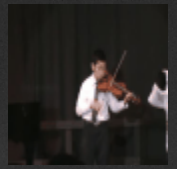
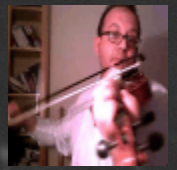



country line dancing

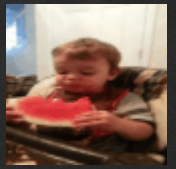
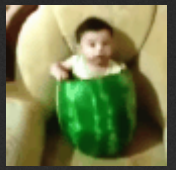





riding or walking with horse



playing violin

eating watermelon

Kinetics

~~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 Release

DeepMind

Kinetics

About

Kinetics is a large-scale, high-quality dataset of YouTube video URLs which include a diverse range of human focused actions. Our aim in releasing the Kinetics dataset is to help the machine learning community to advance models for video understanding.

The dataset consists of approximately 300,000 video clips, and covers 400 human action classes with at least 400 video clips for each action class. Each clip lasts around 10s and is labeled with a single class. All of the clips have been through multiple rounds of human annotation, and each is taken from a unique YouTube video. The actions cover a broad range of classes including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging.

Kinetics forms the basis of an international human action classification competition being organised by ActivityNet.

ACTIVITYNET
Large Scale Activity Recognition Challenge

CVPR July 21-26 2017

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**

Evaluating Actions in Videos

Does this video clip contain the human action
playing drums?

Instructions

We would like to find videos that contain real humans performing actions e.g. scrubbing their face, jumping, kissing 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
- Video does not play, does not contain a human, is an image, cartoon or a computer game.

Combine, split, and filter classes

DeepMind

Action list

Person Actions (Singular)

e.g. waving, blinking, running, jumping

Person-Person Actions

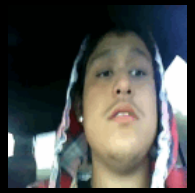
e.g. hugging, kissing, shaking hands

Person-Object Actions

e.g. opening door, mowing lawn, washing dishes

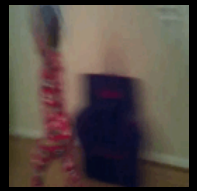
Action list

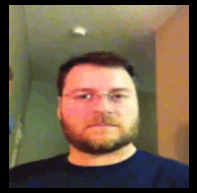
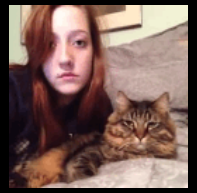
Person Actions (Singular)

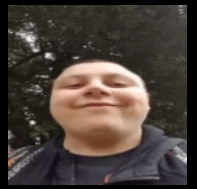
Pumping Fist



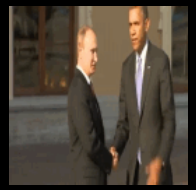



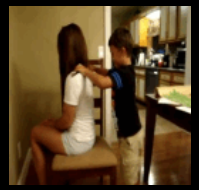

Shaking Head




Action list

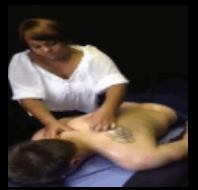
Person-person actions

Shaking Hands


Massaging Back






Making People Feel Welcome on University of Michigan's North Campus

Action list

Person-object actions

Action list

Person-object actions

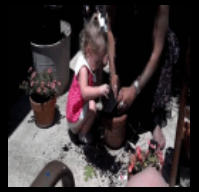
Action list

Person-object actions

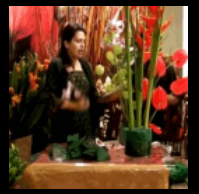



Planting Flowers



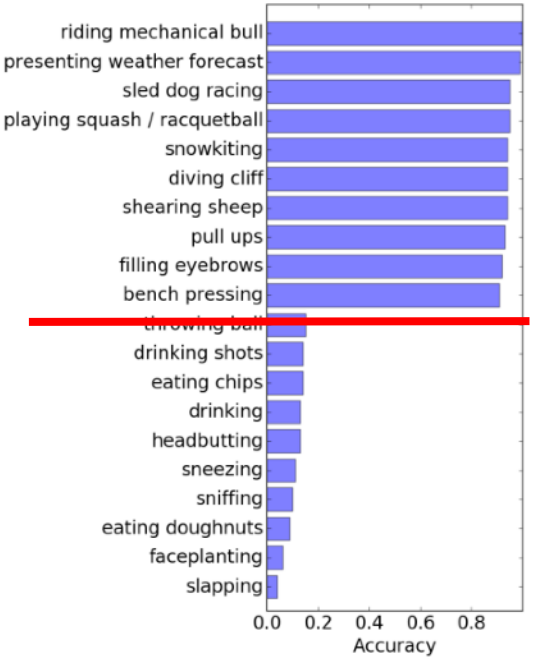






Arranging Flowers



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)

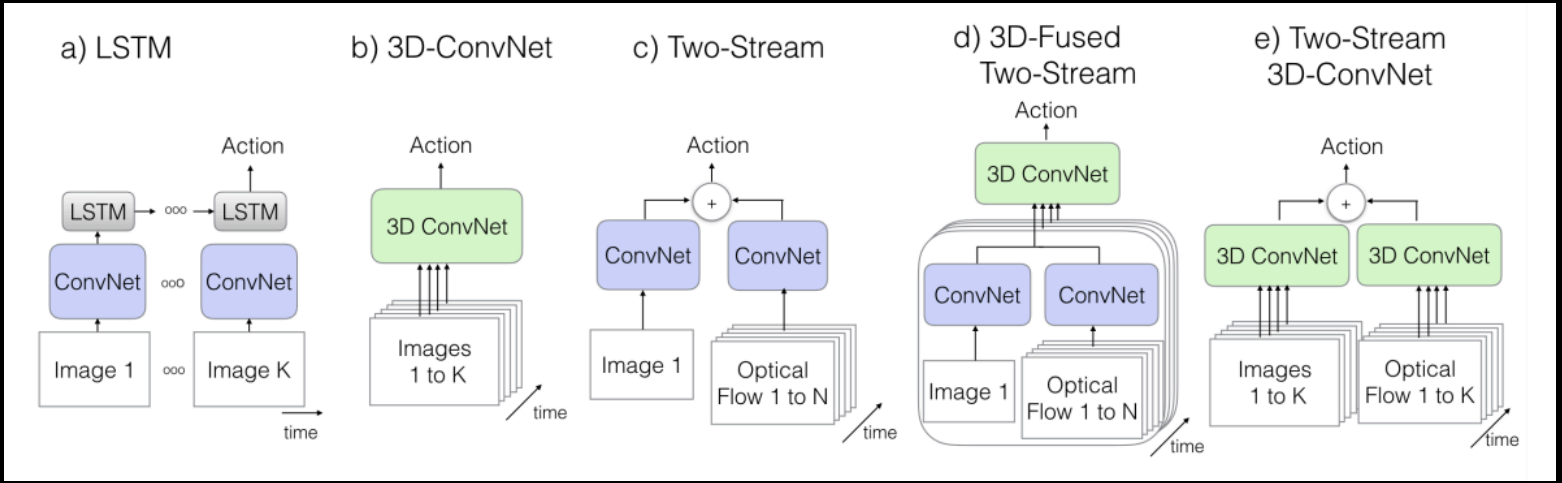
D'oh! You'd never believe it, but in a new research paper, computer scientists at Google DeepMind have admitted that its artificial intelligence technology still struggles to identify many common human behaviours that Homer Simpson exhibits – whether it's eating doughnuts or crisps, falling on his face,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

DeepMind

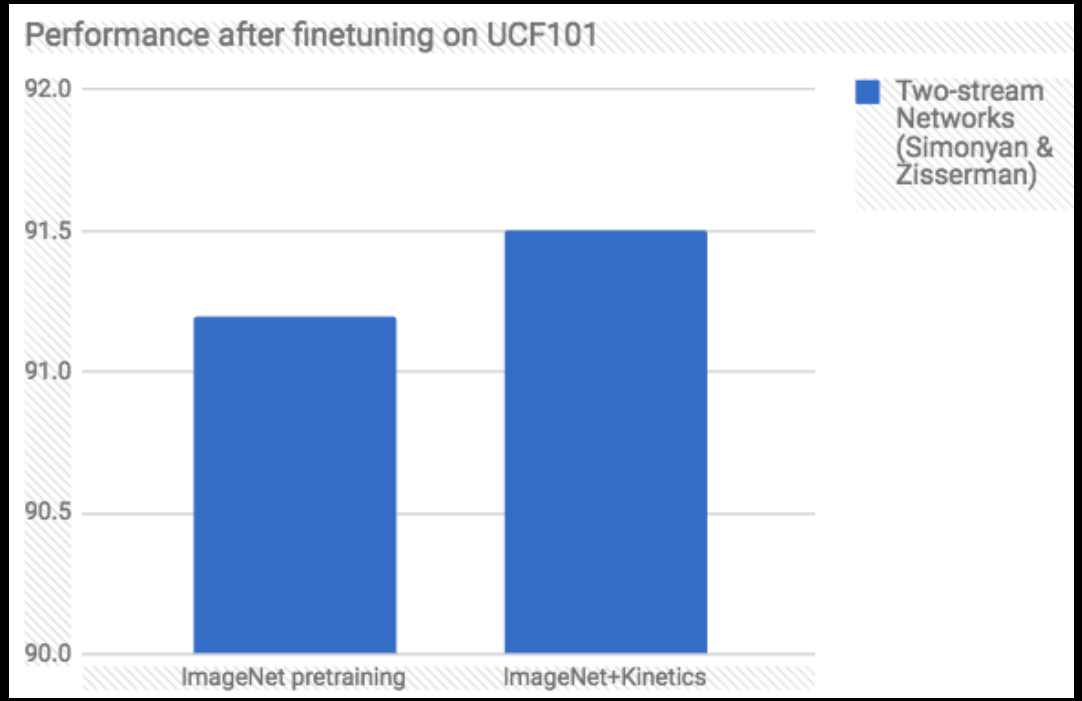
2. The I3D model

Quo Vadis, Action Recognition?
A New Model and the Kinetics Dataset

- Initial hypothesis: given Kinetics, existing models will obliterate UCF-101, HMDB-51, etc.

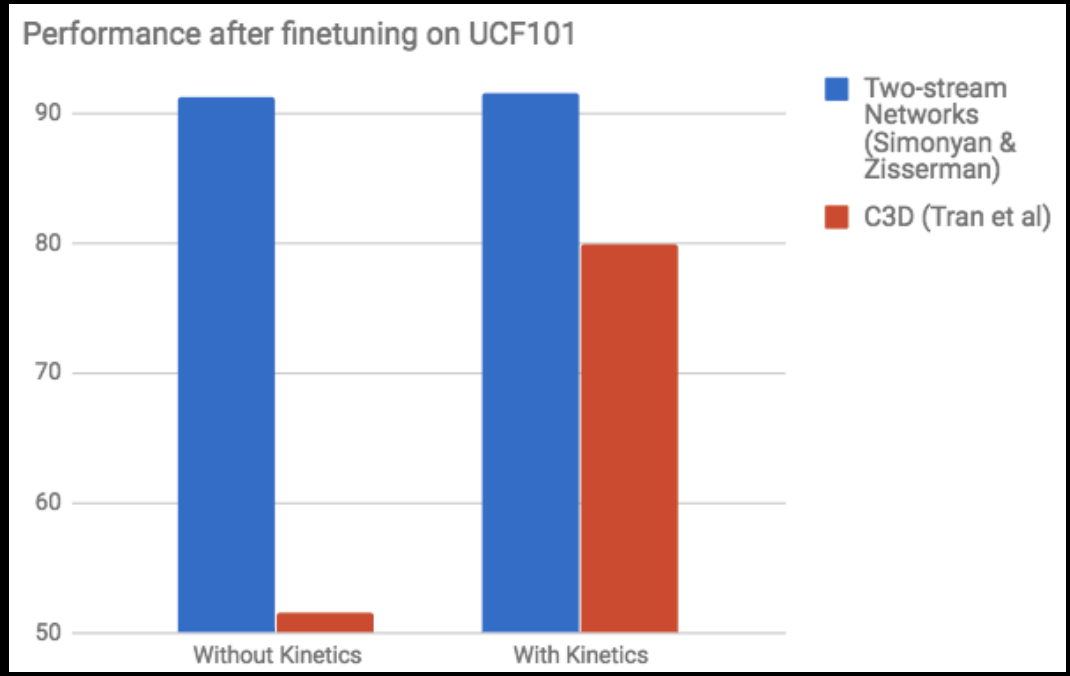


Quo Vadis, Action Recognition?
A New Model and the Kinetics Dataset



Quo Vadis, Action Recognition?
A New Model and the Kinetics Dataset

Spatiotemporal models (3D ConvNets) seem to benefit more from pre-training in videos



C3D

- 8 convolutional layers
- 79M parameters
- Inputs: 112x112, 16-frame clips

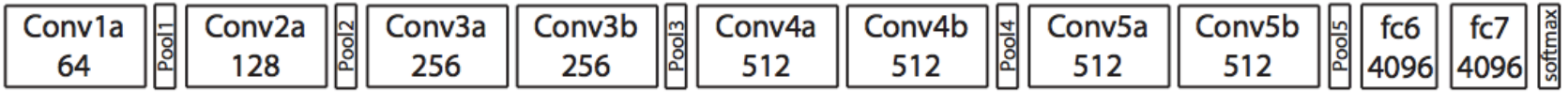


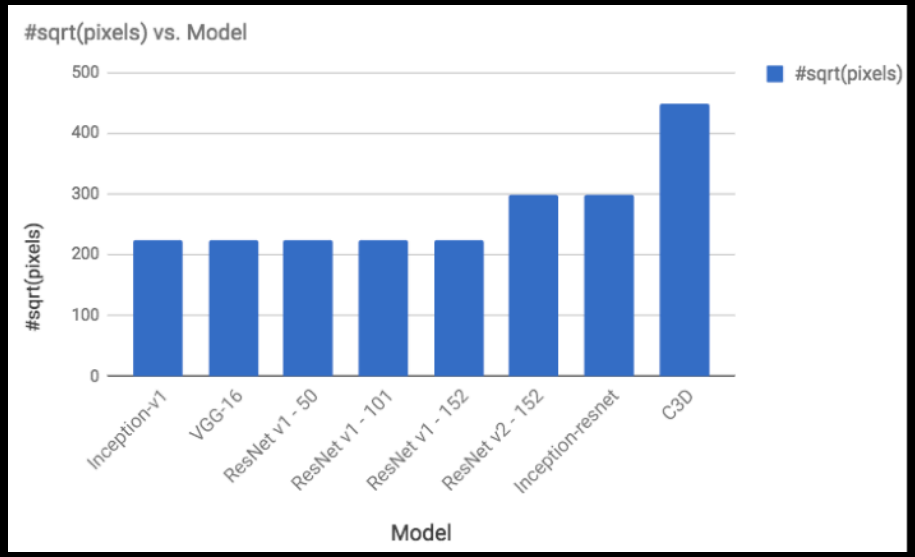
Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling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

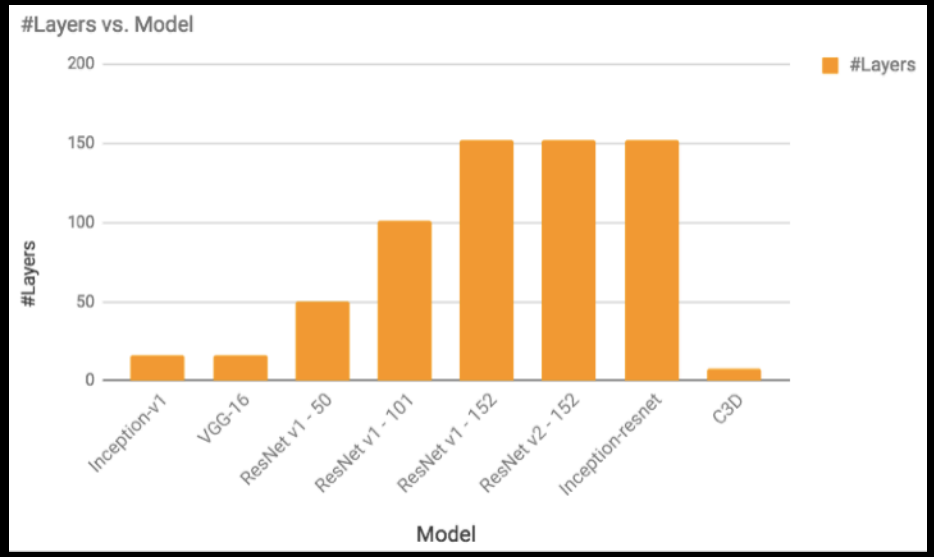
Learning Spatiotemporal Features with 3D Convolutional Networks.
Tran et al, CVPR 2015

Core issue: balance between amount of signal and computation

C3D:

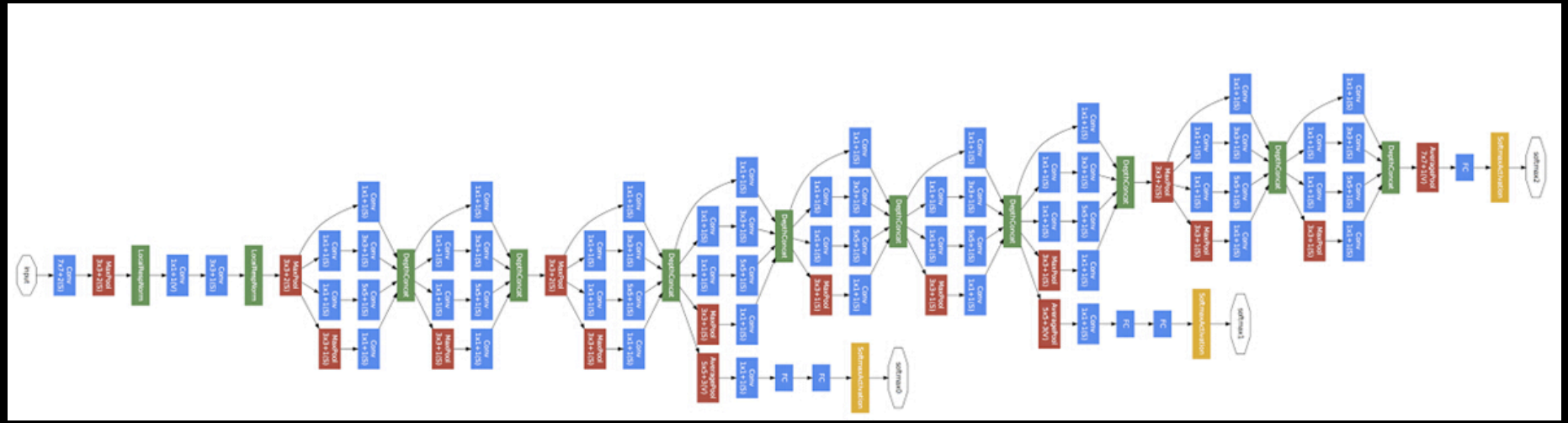
- 8 convolutional layers
- 79M parameters
- Inputs: 112x112, 16-frame clips



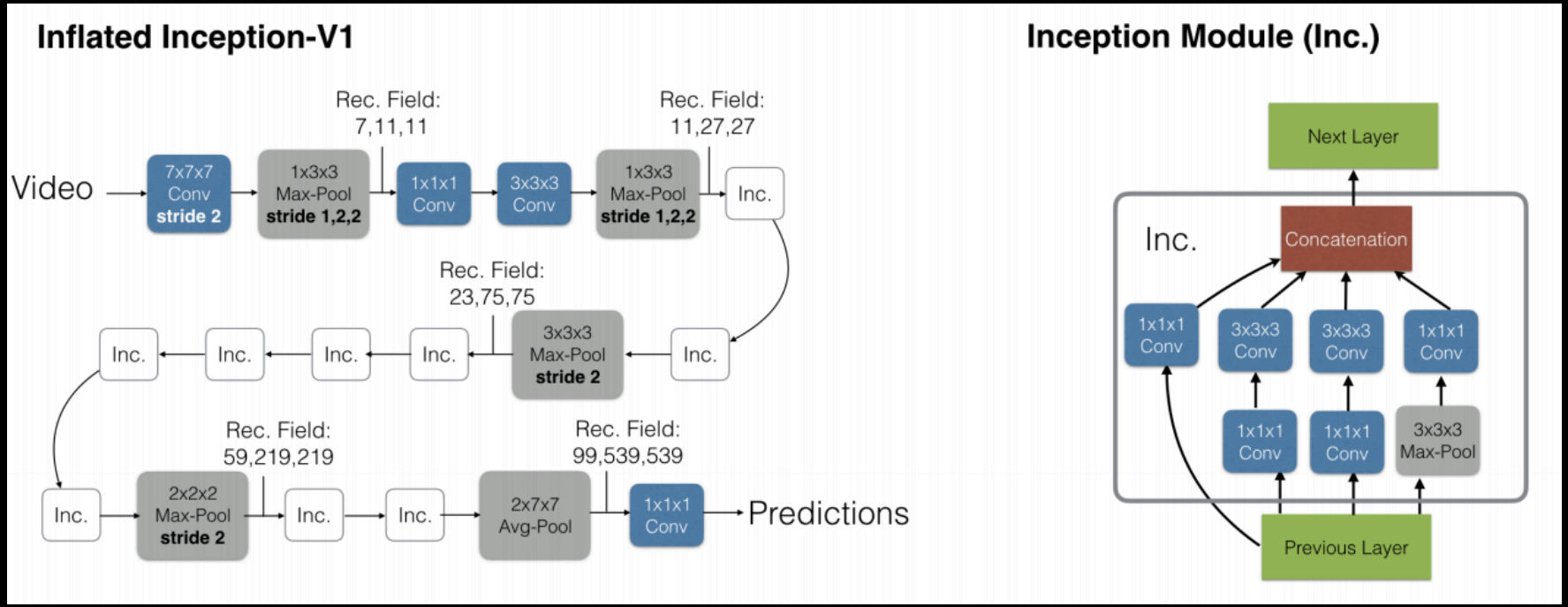


Google's Inception-V1 ImageNet classifier

Going deeper with convolutions, Szegedy et al, CVPR 2015



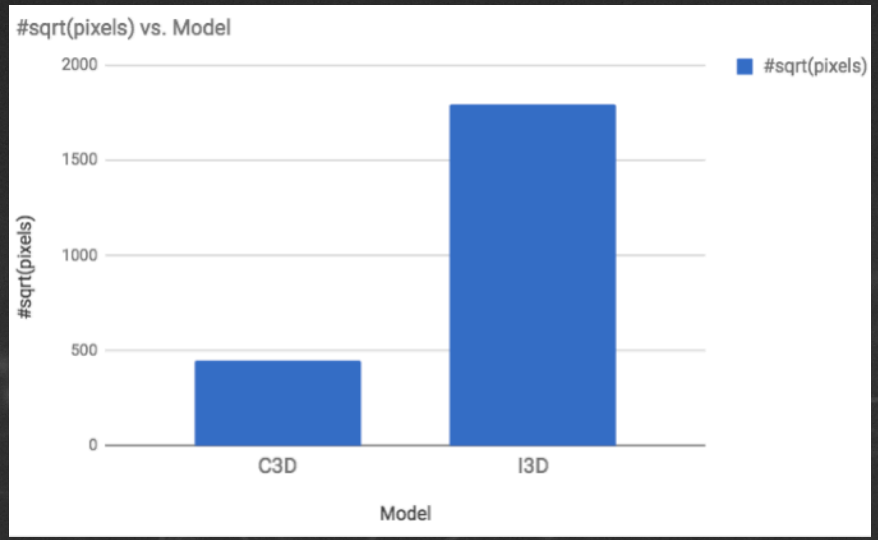
New 3D ConvNet: Inflated 3D Inception (I3D)

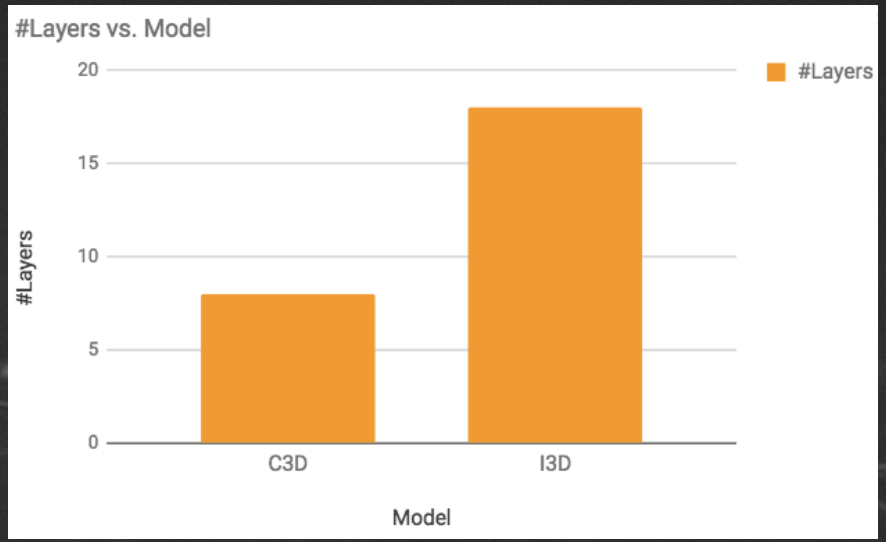


Core issue: balance between amount of signal and computation

C3D:
- 8 convolutional layers
- 79M parameters
- Inputs: 112x112, 16-frame clips

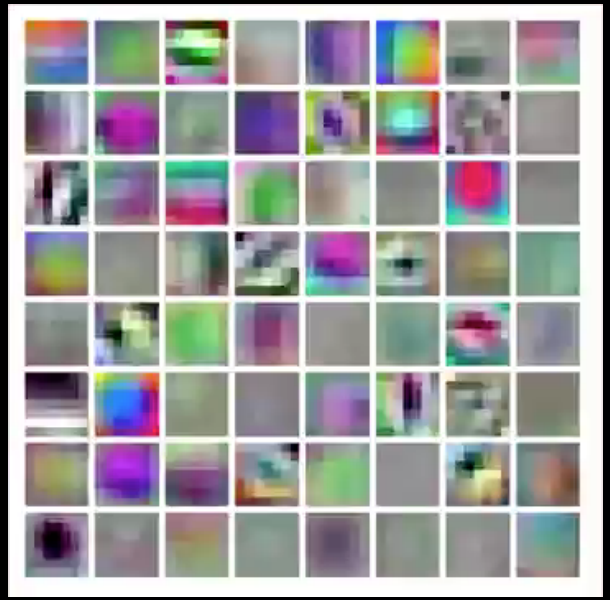
I3D:
- 18 convolutional layers
- 12M parameters
- Inputs: 224x224, 64-frame clips



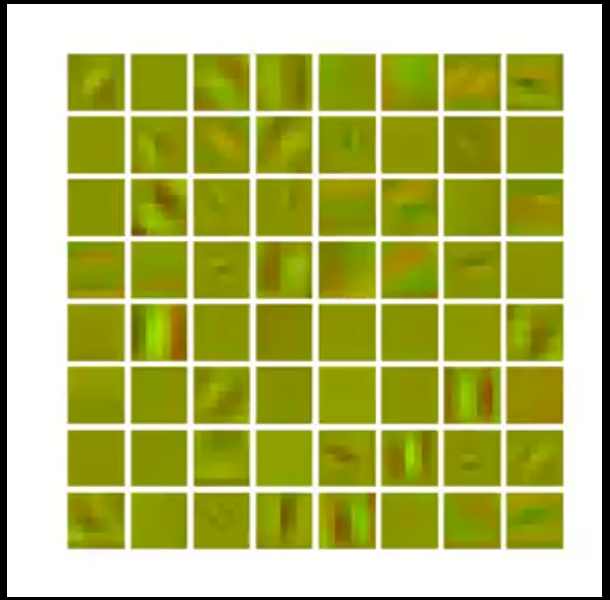


I3D Conv1 filters, trained in Kinetics

RGB

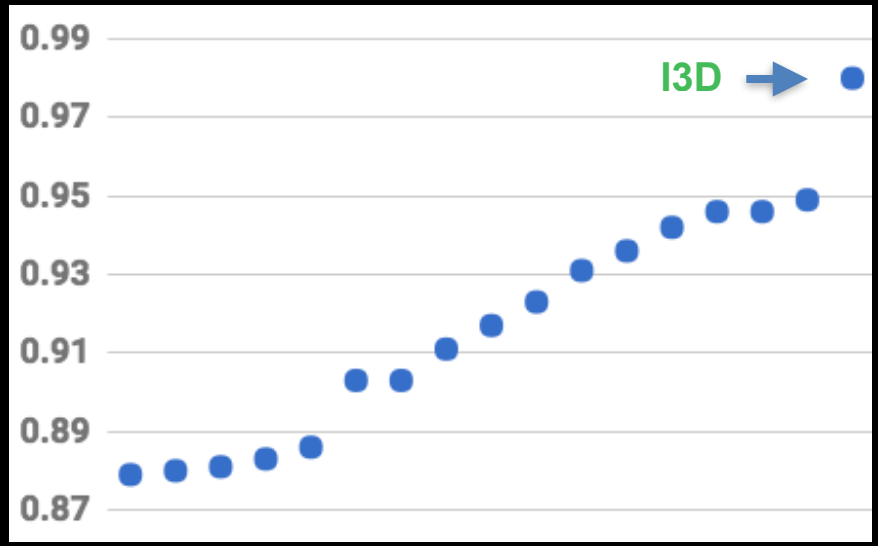


Flow

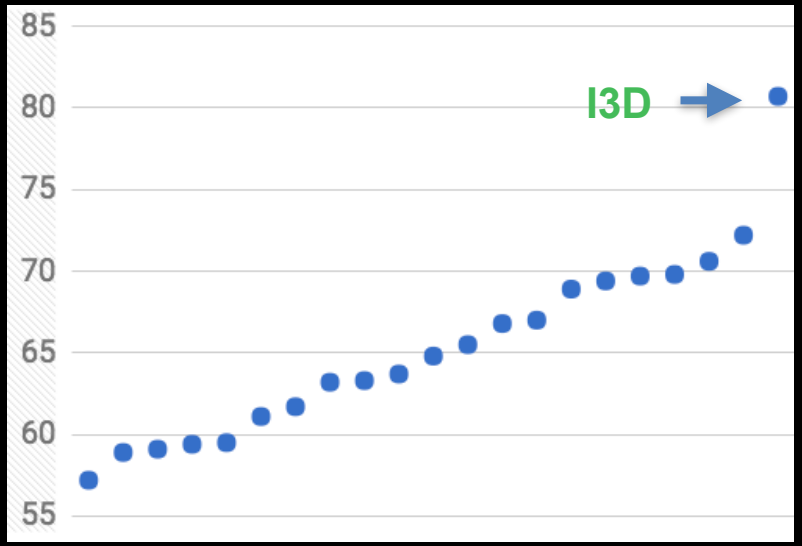


I3D-Kinetics transfer performance (two stream, flow+rgb)

UCF-101

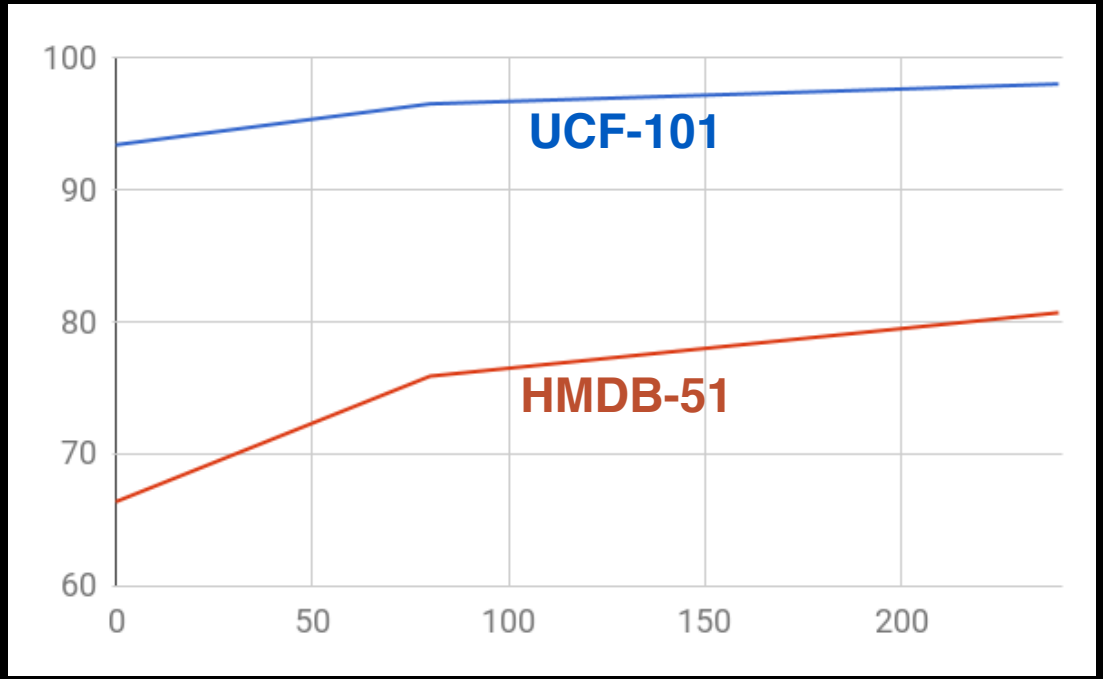


HMDB-51



Kinetics pre-training, comparison with state-of-the-art (compilation of results from actionrecognition.net)

Performance as function of # Kinetics examples



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

Performance on Kinetics

- Final score: 17.2 (would be tied for **5th** in the challenge)

Model	ImageNet + Kinetics	Kinetics
RGB-I3D,	71.1 / 89.3	68.4 / 88.0
Flow-I3D,	63.4 / 84.9	61.5 / 83.4
Two-Stream I3D	74.2 / 91.3	71.6 / 90.0

Guess the action (**hint:** actions likely to be performed by the lake after the talk)

Opening Bottle (64% correct)

Model Predictions:

Opening Bottle

Opening Bottle

Dancing Gangnam Style

Guess the action

Tasting Beer (71% correct)

Model Predictions:

Tasting Beer

Tasting Beer

Giving or Receiving Award

Guess the action

Dancing Charleston (68% correct)

Model Predictions:

Dancing Charleston

Dancing Charleston

Zumba

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

Pretrained I3D models will be available within 2 weeks at deepmind.com/kinetics

Conclusions

- Maybe a fresh new beginning in video research informed by - but detached - from successes in image understanding
- Just scratching the surface of video modeling on Kinetics
- Will the models generalize to other video tasks ?

Thanks!