Insight
Centre for Data Analytics

DEEP
LEARNING
WORKSHOP

Dublin City University
21-22 May 2018

DCU

Day 2 Lecture 1

Video

Eva Mohedano
eva.mohedano@insight-centre.org

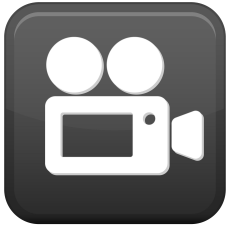
Postdoctoral Researcher
Insight Centre for Data Analytics
Dublin City University

Contents

Contents

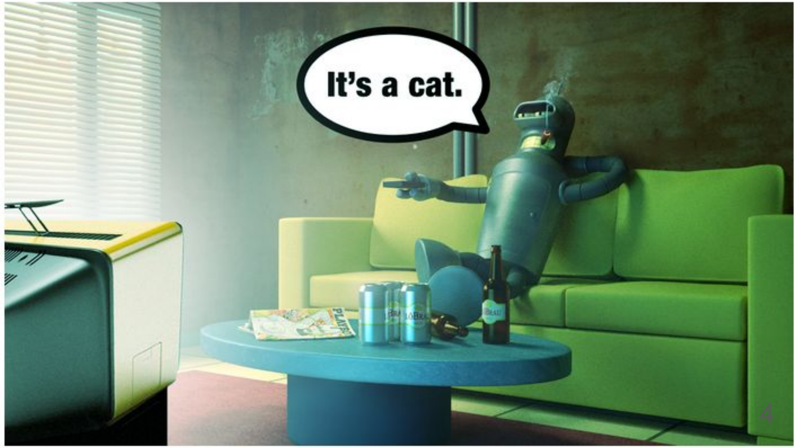
- **Video Classification**
- CNN Architectures
- Comments & thoughts
- Conclusions

Video Classification



Deep Learning Model

Video Content Label(s)



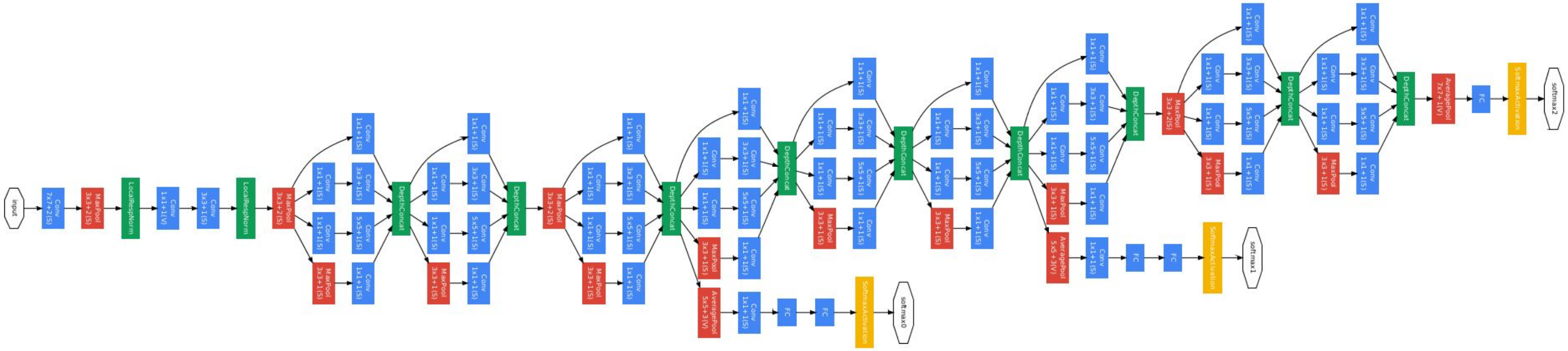
What is a Video?

- Formally, a video is a 3D signal
  - Spatial coordinates: x, y
  - Temporal coordinate: t

- If we fix t, we obtain an image. We can understand videos as sequences of images (a.k.a. frames)

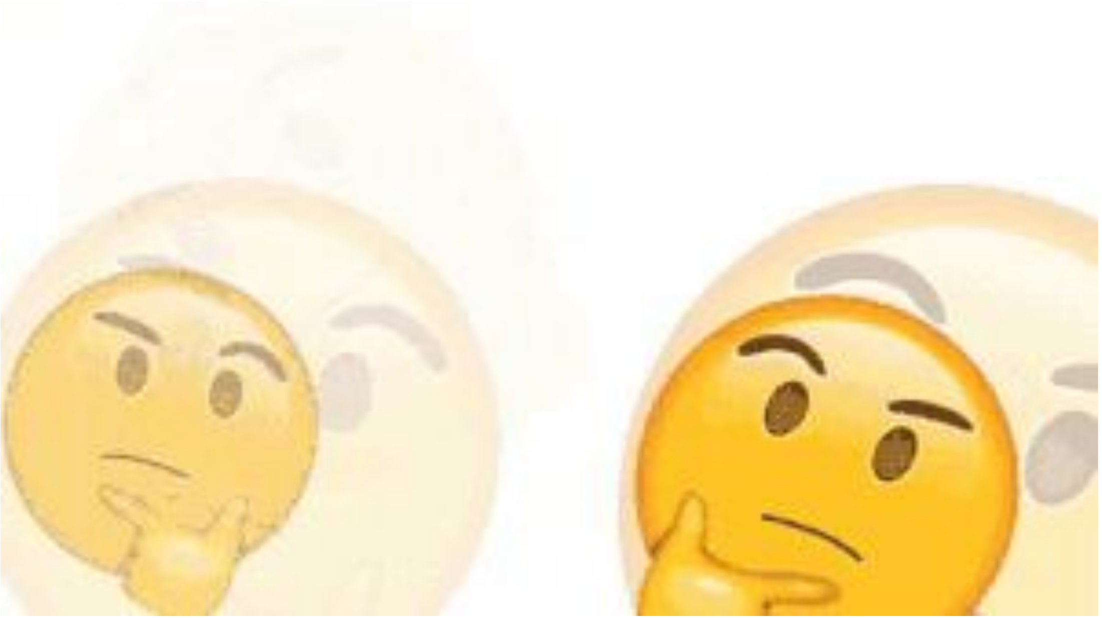
What do we do with Images?

Convolutional Neural Networks (CNNs) which provides the state of the art in still image analysis



What do we do with Videos?

How to extend CNNs to work with image sequences?



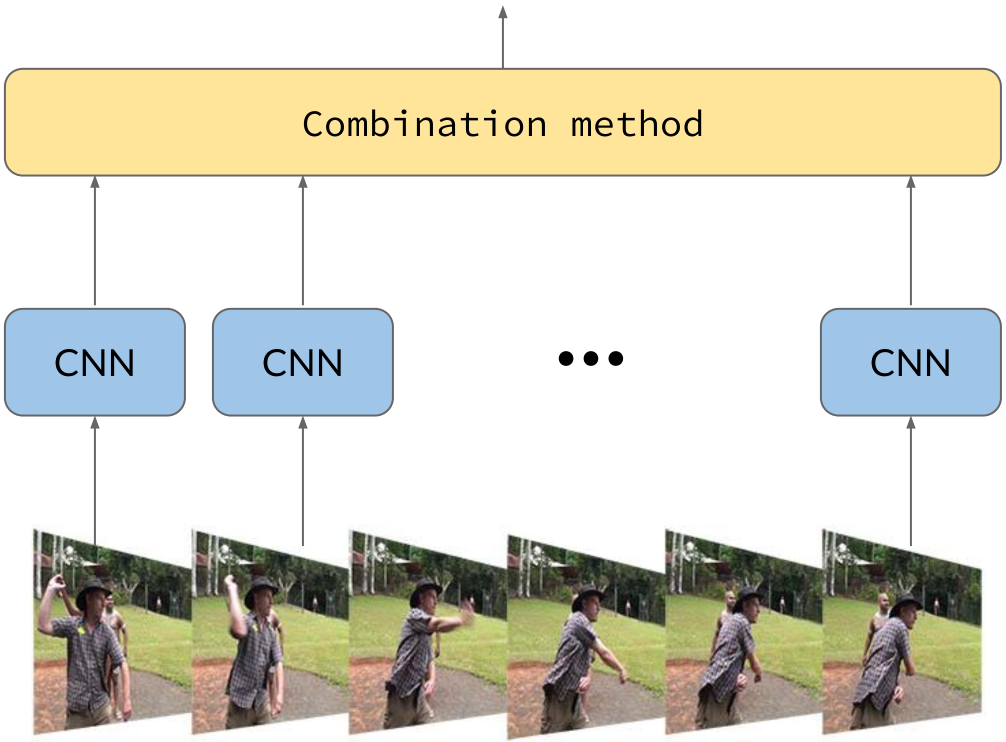
Contents

- Video Classification
- **CNN Architectures**
- Comments & thoughts
- Conclusions

CNN Architectures for Video

1. ConvNet+Pooling
2. ConvNet+RNN
3. 3D Convolutional models
4. Two Stream CNNs
5. Two Streams 3D-CNNs

ConvNet+Pooling

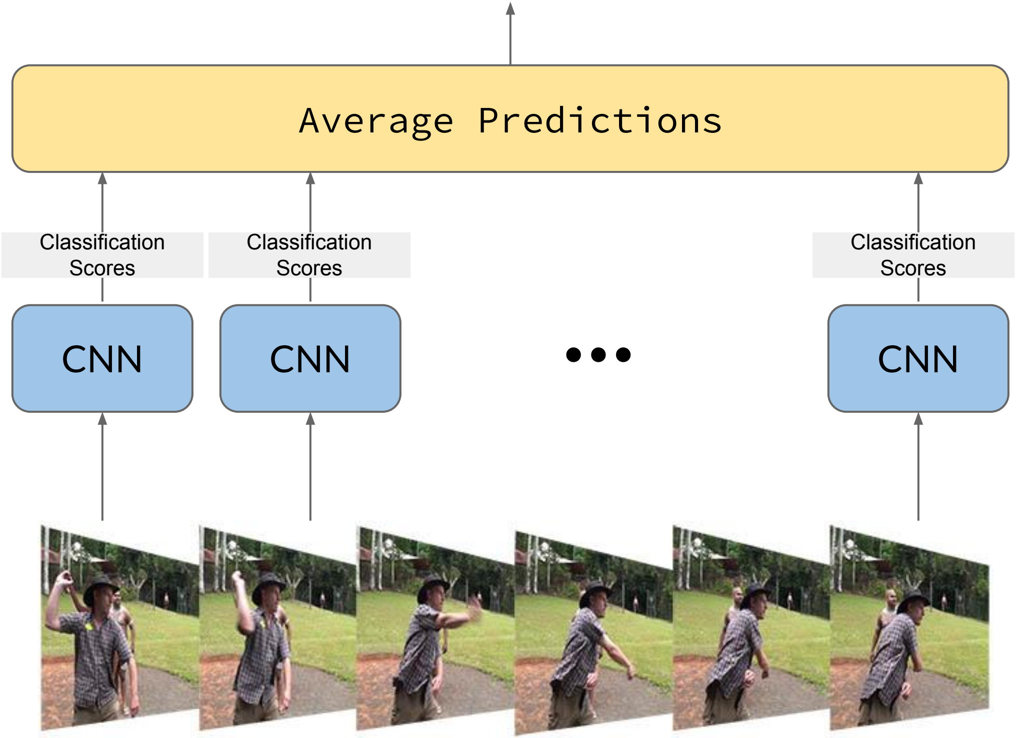


- Combination is commonly implemented as a small NN on top of a **pooling operation** (i.e max, sum, average, BoW, VLAD)

Problem: Pooling is not aware of temporal order!
(Sometimes this is not a problem. Y8M)

- Combination is implemented as a small NN on top of a **Recurrent Neural Network**

Single-frame model

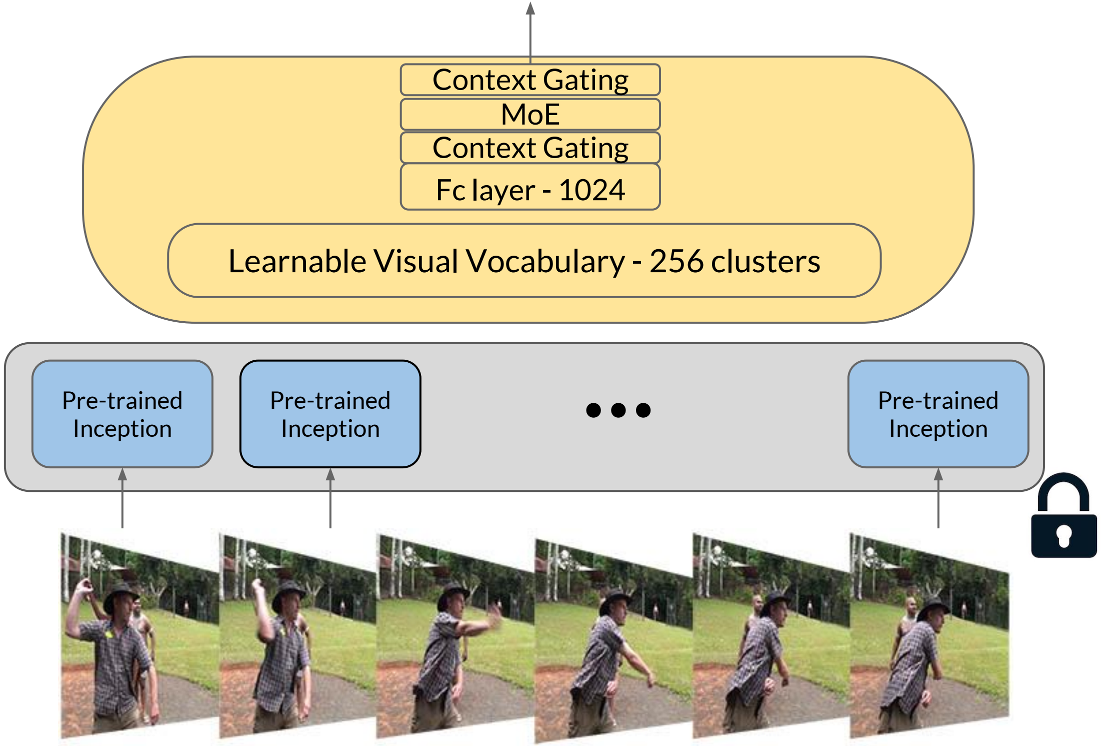


The 'super simple' approach
(Worth to try as a baseline)

- Take pre-trained CNN
- Remove the last classifier layer
- Plug new layer to classify your classes
- Train new layer and/or fine-tune early CNN layers
- Average predictions across frames
- Baseline set :-)

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. Large-scale video classification with convolutional neural networks. CVPR 2014

ConvNet+Pooling



NetVLAD pooling
(State of the Art in Youtube8M)

NetVLAD Module

$$VLAD(j,k) = \sum_{i=1}^{N} a_k(x_i)(x_i(j) - c_k(j)),$$

$$a_k(x_i) = \frac{e^{w_k^T x_i + b_k}}{\sum_{j=1}^{K} e^{w_j^T x_i + b_j}}$$

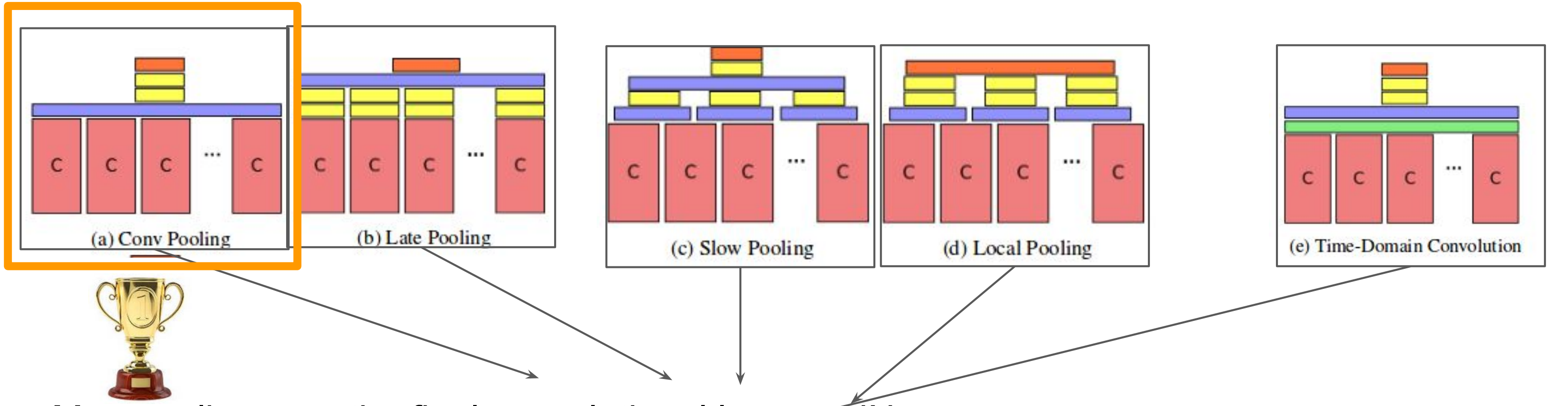
Context Gating Module

$$Y = \sigma(WX + b) \circ X,$$

Learnable pooling with context gating for video classification - A. Miech, I. Laptev and J. Sivic arXiv2017

ConvNet+Pooling

Analysis of different feature-pooling architectures



Max-pooling over the final convolutional layer

Hierarchically combines frame-level information from smaller temporal windows. Two-stage pooling strategy.

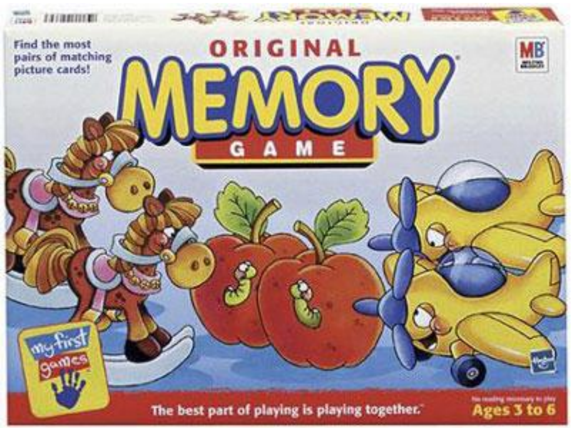
Small temporal windows with NN + final fc pooling

Time-domain convolution (3D conv) prior to max-pooling

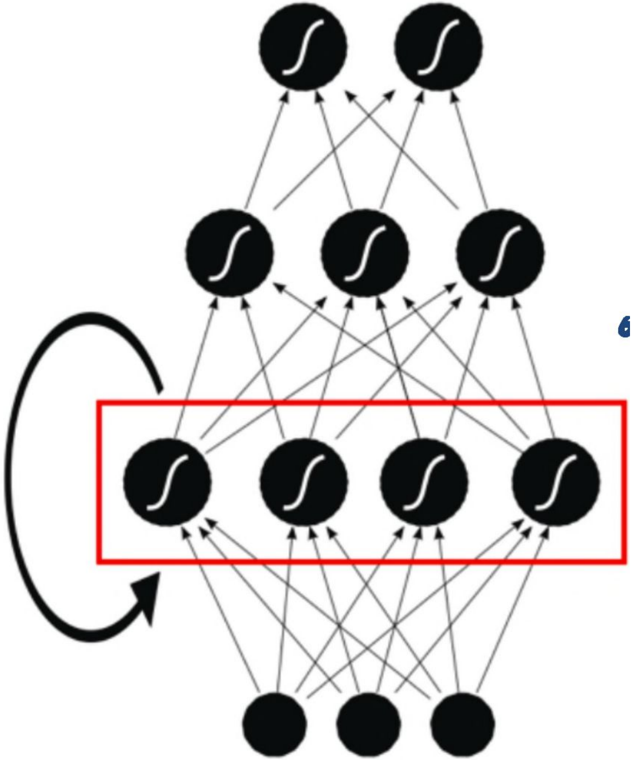
CNN Architectures for Video

1. ConvNet+Pooling
2. **ConvNet+RNN**
3. 3D Convolutional models
4. Two Stream CNNs
5. Two Streams 3D-CNNs

Recurrent Neural Networks

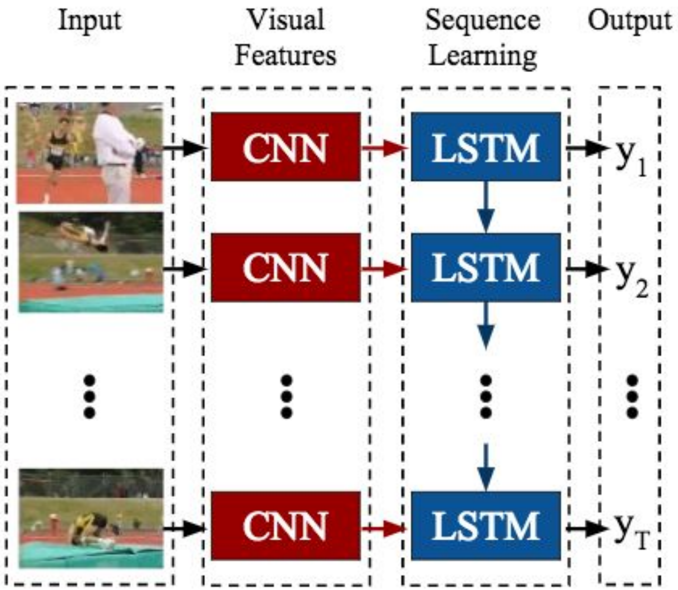


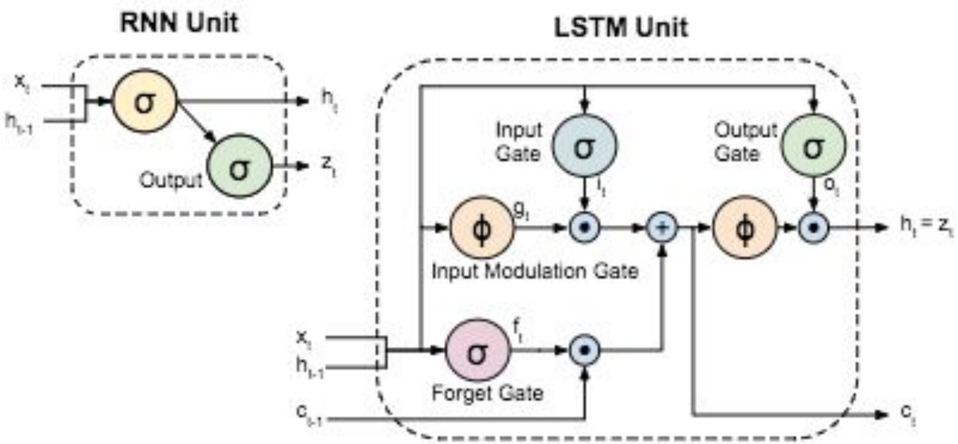
The hidden layers and the output depend from previous states of the hidden layers



CNN+LSTM

Training on short clips of 30frames (1fps) -- 100 frame videos on UCF101
Inference overlapped clips + average prediction



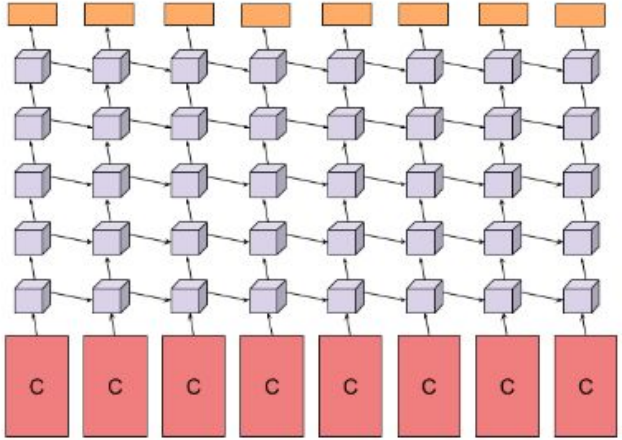


Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

CNN+RNN

Recurrent NN model

LSTM architecture



Performance on Sports-1M (AlexNet)

Method	Clip Hit@1	Hit@1	Hit@5
Conv Pooling	**68.7**	**71.1**	**89.3**
Late Pooling	65.1	67.5	87.2
Slow Pooling	67.1	69.7	88.4
Local Pooling	68.1	70.4	88.9
Time-Domain Convolution	64.2	67.2	87.2

Performance on Sports-1M

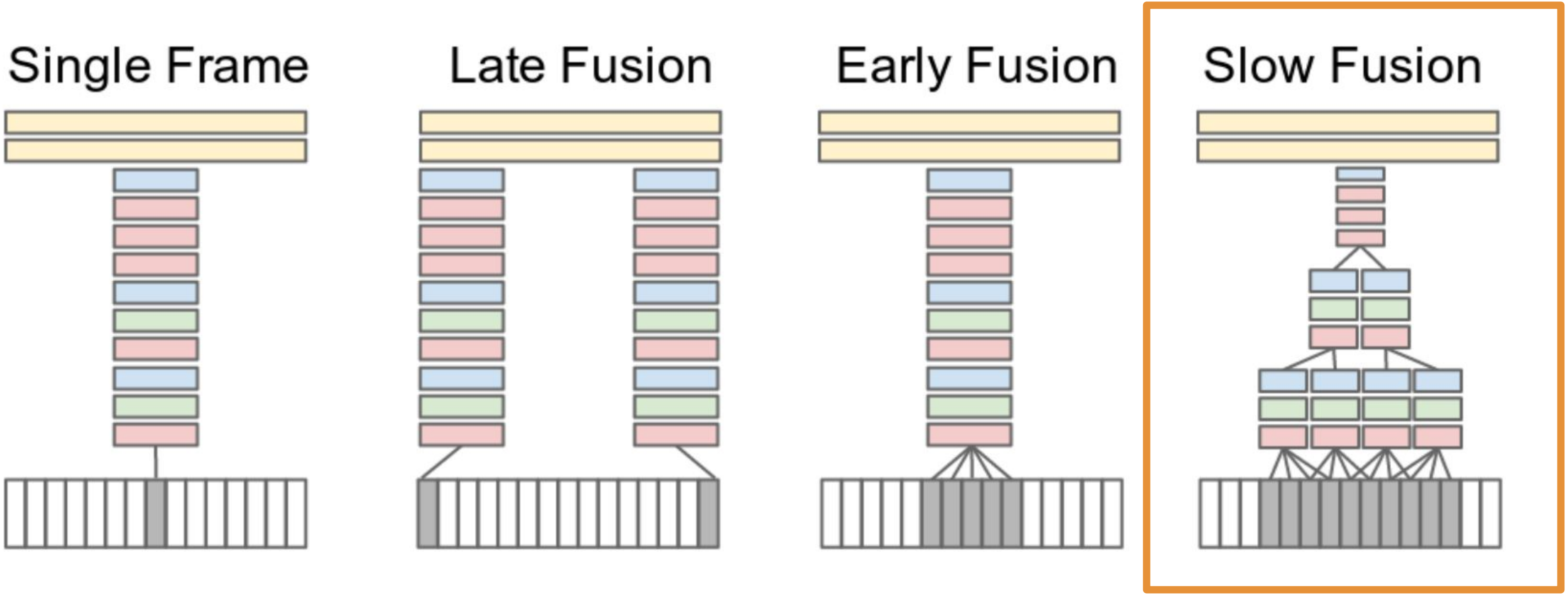
Method	Hit@1	Hit@5
AlexNet single frame	63.6	84.7
GoogLeNet single frame	**64.9**	**86.6**
LSTM + AlexNet (fc)	62.7	83.6
LSTM + GoogLeNet (fc)	**67.5**	**87.1**
Conv pooling + AlexNet	70.4	89.0
Conv pooling + GoogLeNet	**71.7**	**90.4**

Beyond Short Snippets: Deep Networks for Video Classification - Joe Yue-Hei Ng et al. CVPR 2015

CNN Architectures for Video

1. ConvNet+Pooling
2. ConvNet+RNN
3. **3D Convolutional models**
4. Two Stream CNNs
5. Two Streams 3D-CNNs

Multi-frame CNNs

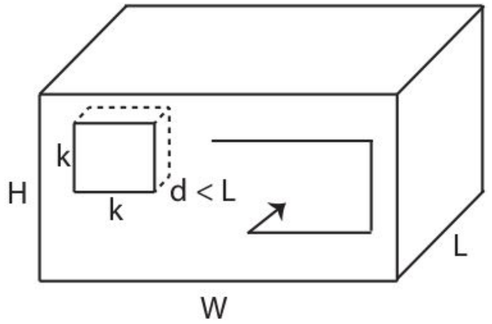


Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. . Large-scale video classification with convolutional neural networks. CVPR 2014

3D-CNN

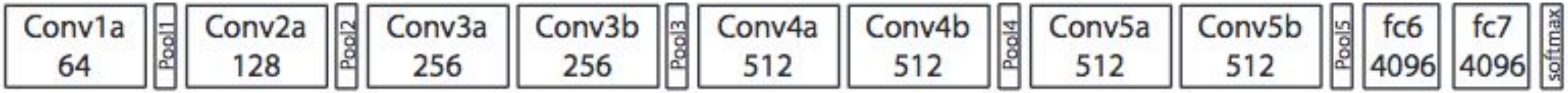
We can add an extra dimension to standard CNNs:

- An image is a HxWxD tensor: MxNxD’ conv filters
- A video is a TxHxWxD tensor: KxMxNxD’ conv filters



Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3D convolutional networks" ICCV 2015

3D-CNN



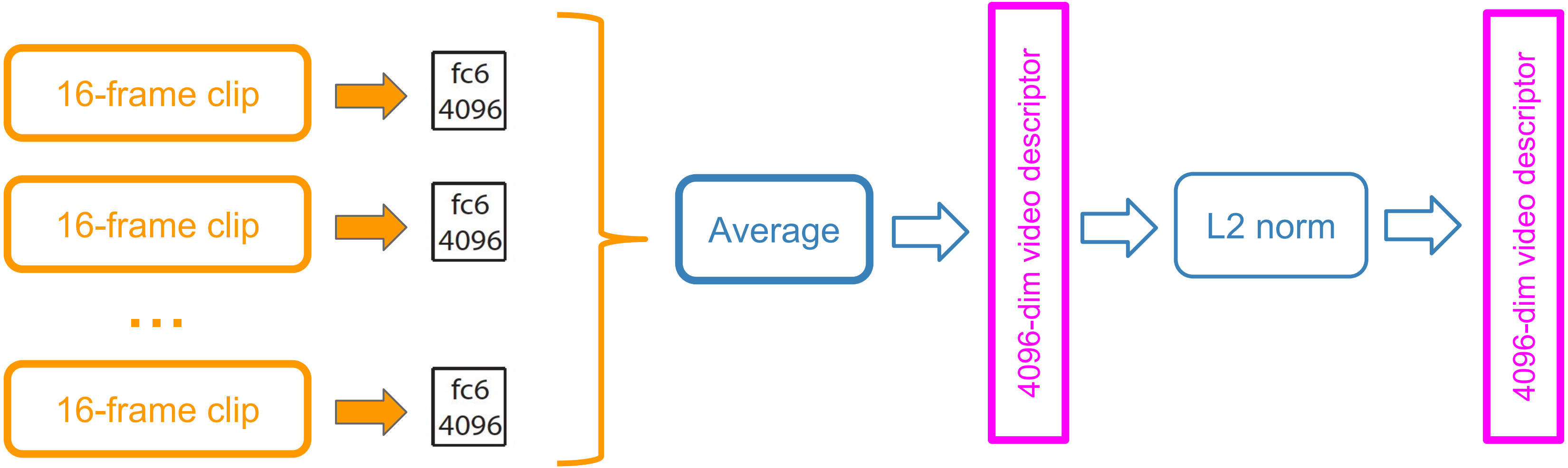
8 conv, 5 max-pooling, 3 fully connected, softmax output layer
3D kernels: 3 x 3 x 3 with stride 1
Pool layers 2 x 2 x 1 (first layer), 2 x 2 x 2 (the rest)

We preserve the spatio-temporal information across the layers
Hard to train
Hard to explicitly learn local temporal feature relations

Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "[Learning spatiotemporal features with 3D convolutional networks]" ICCV 2015

3D-CNN

The video needs to be split into chunks (also known as *clips*) with a number of frames that fits the receptive field of the C3D. Usually clips have 16 frames.



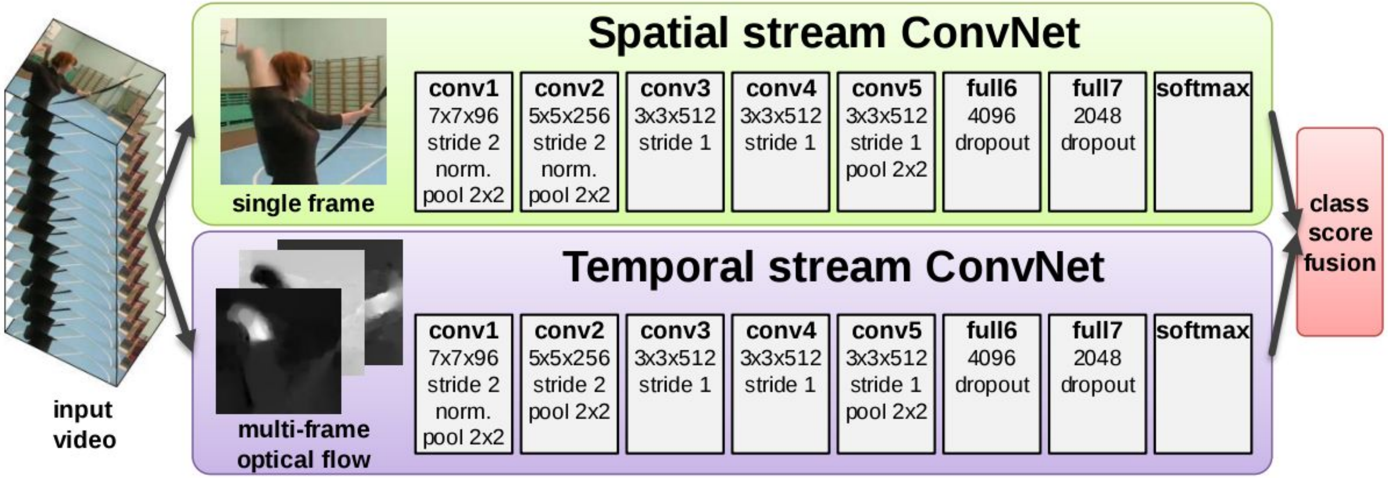
Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "[Learning spatiotemporal features with 3D convolutional networks]" ICCV 2015

CNN Architectures for Video

1. ConvNet+Pooling
2. ConvNet+RNN
3. 3D Convolutional models
4. **Two Stream CNNs**
5. Two Streams 3D-CNNs

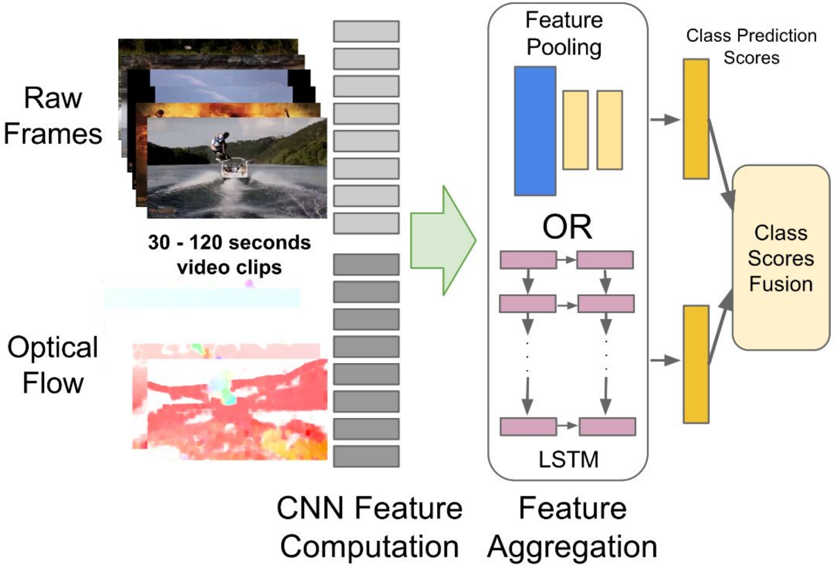
Two Stream Networks

Problem: Single frame models do not take into account motion in videos.
Solution: extract optical flow for a stack of frames and use it as an input to a CNN.



Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." NIPS 2014.

Two Stream Networks

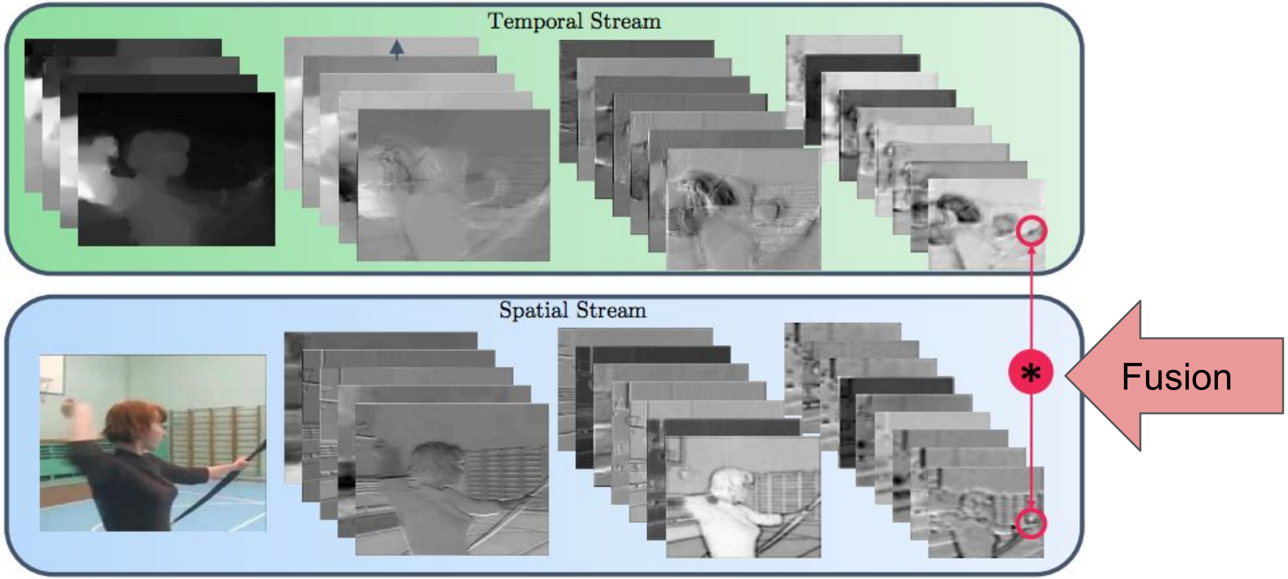


Method	Hit@1	Hit@5
LSTM on Optical Flow	59.7	81.4
LSTM on Raw Frames	72.1	90.6
LSTM on Raw Frames + LSTM on Optical Flow	73.1	90.5
30 frame Optical Flow	44.5	70.4
Conv Pooling on Raw Frames	71.7	90.4
Conv Pooling on Raw Frames + Conv Pooling on Optical Flow	71.8	90.4

Optical flow provide the fine-grained temporal detail (15fps optical flow vs 1fps RGB frame extraction)

Two Stream CNN Networks

Modality fusion (appearance and temporal) at conv layer. → 3Dconv layer + Pooling instead of late fusion (average score)



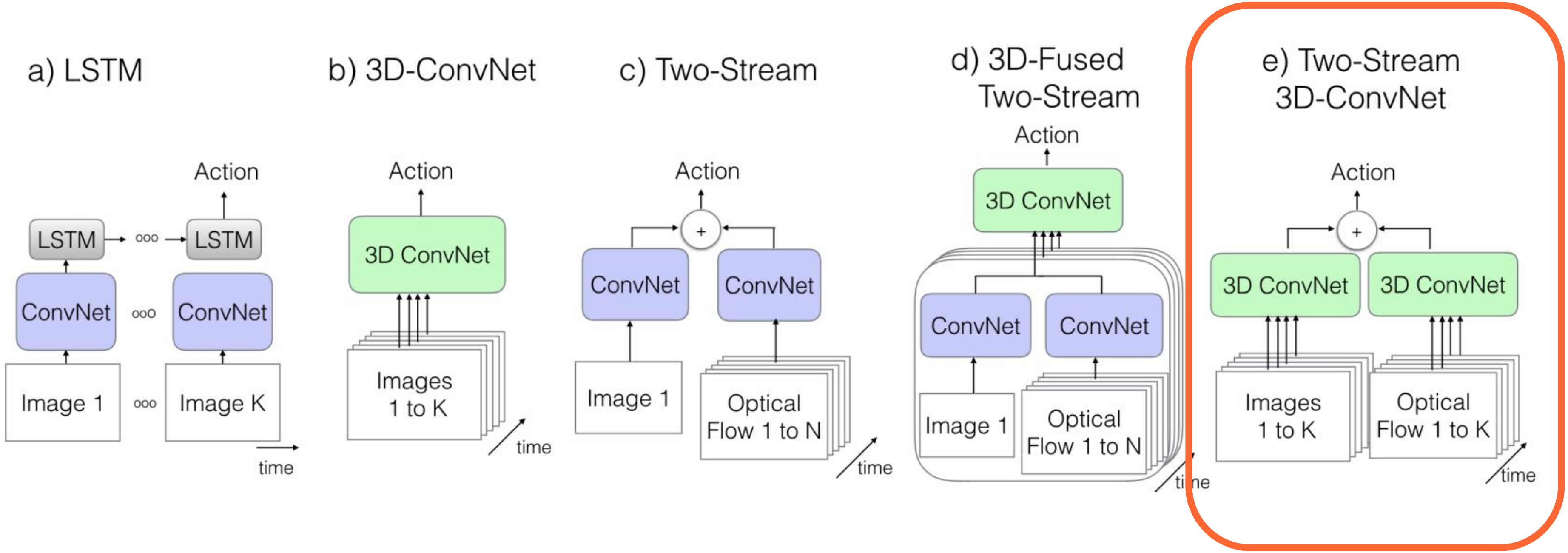
Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. ["Convolutional two-stream network fusion for video action recognition."]() CVPR 2016. [[code]]()

CNN Architectures for Video

1. ConvNet+Pooling
2. ConvNet+RNN
3. 3D Convolutional models
4. Two Stream CNNs
5. **Two Stream 3D-CNNs**

Two Streams 3D-CNN

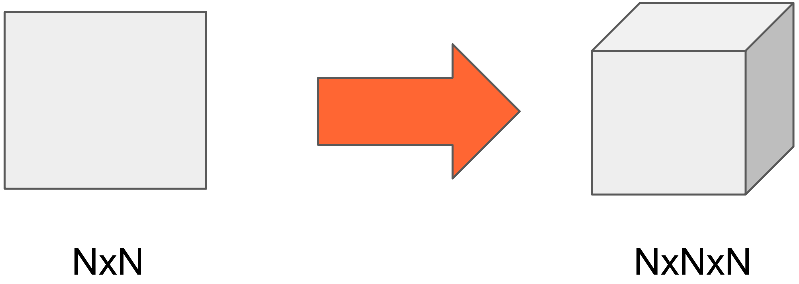
Kinetics Dataset for training



Carreira, J., & Zisserman, A. . Quo vadis, action recognition? A new model and the kinetics dataset. CVPR 2017. [code]

Two Streams 3D-CNN

Adapt 2D CNNs found for ImageNet classification to 3D convolutions

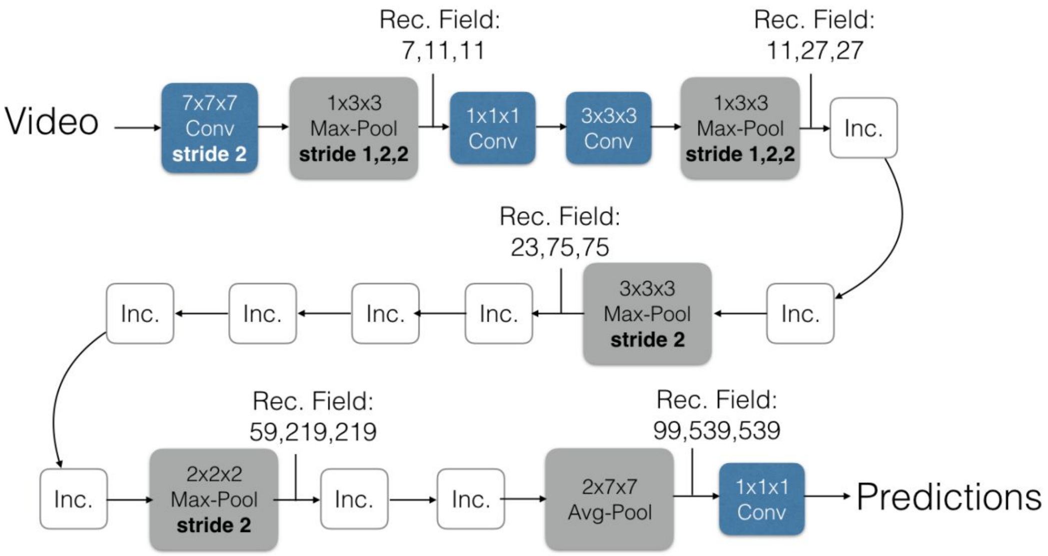


3D models are initialized with ImageNet images transformed into ‘boring’ video sequences.

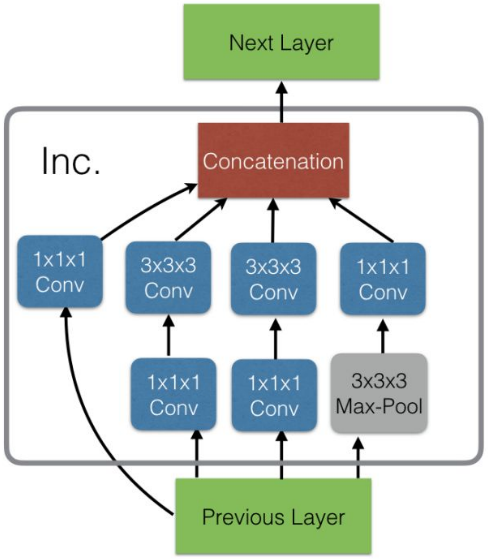
Carreira, J., & Zisserman, A. . [Quo vadis, action recognition? A new model and the kinetics dataset](). CVPR 2017. [code]

Two Streams 3D-CNN

Inflated Inception-V1



Inception Module (Inc.)



Carreira, J., & Zisserman, A. . [Quo vadis, action recognition? A new model and the kinetics dataset]. CVPR 2017. [[code]]

Two Streams 3D-CNN

Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**74.1**	**69.6**	**78.7**

- 25fps RGB stream
- Two Stream I3D trained on 64 GPUs!!!!

Carreira, J., & Zisserman, A. . Quo vadis, action recognition? A new model and the kinetics dataset. CVPR 2017. [code]

Contents

- Video Classification
- CNN Architectures
- **Comments & thoughts**
- Conclusions

Data Augmentation

- Large amount of parameters in the model

Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet	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.

Carreira, J., & Zisserman, A. . [Quo vadis, action recognition? A new model and the kinetics dataset](). CVPR 2017. [[code]]()

Large Scale Dataset

Large scale datasets (Youtube 8M, Kinetics, Sports1M)

- The reference dataset for image classification, **ImageNet, has ~1.3M images**
  - Training a state of the art CNN can take up to **1 weeks** on a single GPU
- Now imagine that we have an ‘ImageNet’ of **1.3M videos**
  - Assuming videos of 30s at 24fps, we have **936M frames**
  - This is 720x ImageNet!
- Videos exhibit a large redundancy in time
  - We can reduce the frame rate without losing too much information

Design a reduced/controlled development dataset, and scale (if possible) to the large one

- Careful with overfitting (too many model parameters)
- Careful with simplifying too much the problem (consiferent a few classes)

Smart frame extraction

- Consecutive frames (25fps) might contain redundant information

- Uniform sampling might not be the best idea
  - Video summarisation techniques?
    - Keyframes + optical flow for temporal information
  - Shot detection
  - I-frames from video codecs

How to deal with really long videos?

Videos might include more than one concept/action related on time

A. Montes, Salvador, A., Pascual-deLaPuente, S., and Giró-i-Nieto, X., "**Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks**", *NIPS Workshop 2016* **(best poster award)**

What about other fusion modalities?

- Optical flow
- Audio
- Text
- ?

Conclusions

- Reviewed some of the popular architectures for video classification
- Two-stream network RGB + Optical flow
- (more computationally expensive) Two-stream 3D convolutional networks
- Pre-trained on large-scale datasets (ideally video s.a Kinetics) key step
- Include more modalities (multiple stream networks)

Thank you!