

#DLUPC

Computer Vision 10

Neural Architectures for Video



Xavier Giro-i-Nieto

xavier.giro@upc.edu

Associate Professor

Universitat Politècnica de Catalunya
Barcelona Supercomputing Center



Acknowledgements



Víctor Campos



Amaia Salvador



Alberto Montes



Santiago Pascual



UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Video lecture



The slide is titled "DEEP LEARNING FOR COMPUTER VISION" and is for "Day 3 Lectures 1 & 2 Video Analysis". It features a Twitter icon and the hashtag #DLUPC. The slide includes a list of instructors, logos for organizers (vilynx, Google Cloud Platform) and supporters (vilynx, Google Cloud Platform), and a small portrait of Víctor Campos. The URL <http://bit.ly/dlcv2018> is also present.

**DEEP LEARNING
FOR COMPUTER VISION**

Summer School at UPC, Telecampus Barcelona, June 28-July 6, 2018

Day 3 Lectures 1 & 2
Video Analysis

#DLUPC

Instructors

Organized by

Supported by

<http://bit.ly/dlcv2018>

Víctor Campos
victor.campos@bsc.es

PhD Candidate
Barcelona Supercomputing Center

BSC



Víctor Campos

victor.campos@bsc.es

PhD Candidate

Barcelona Supercomputing Center



**UNIVERSITAT POLITÈCNICA
DE CATALUNYA**

[Deep Learning for Computer Vision](http://bit.ly/dlcv2018) (UPC 2018)

Outline

- 1. Architectures**
2. Tips and tricks



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

Motivation



track cycling
cycling
track cycling
road bicycle racing
marathon
ultramarathon



ultramarathon
ultramarathon
half marathon
running
marathon
inline speed skating



heptathlon
heptathlon
decathlon
hurdles
pentathlon
sprint (running)



bikejoring
mushing
bikejoring
harness racing
skijoring
carting

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

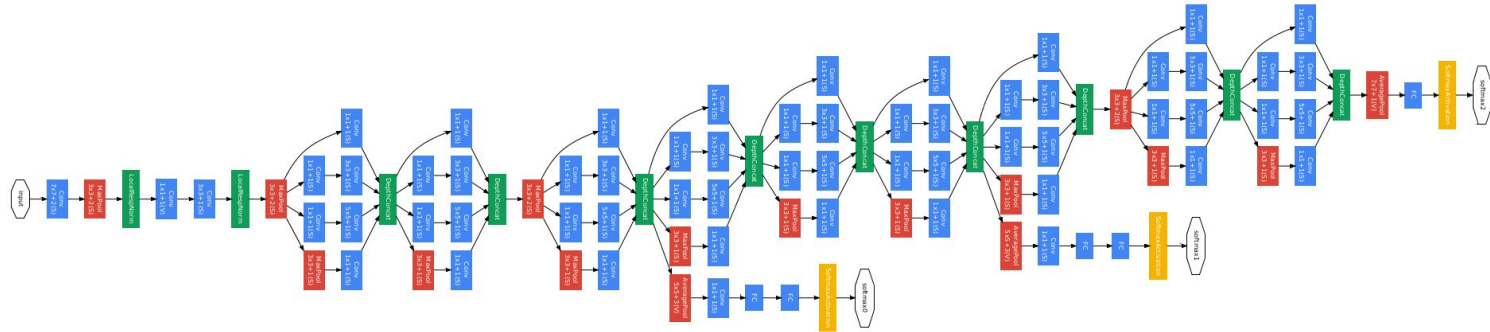
What is a video?

- 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)



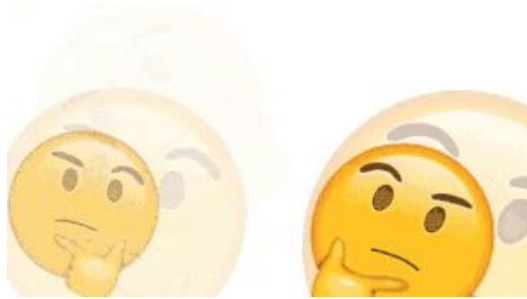
How do we work with images?

Convolutional Neural Networks (CNN) provide state of the art performance on image analysis tasks



How do we work with videos ?

How can we extend CNNs to image sequences?



Deep Video Architectures

Basic deep architectures for video:

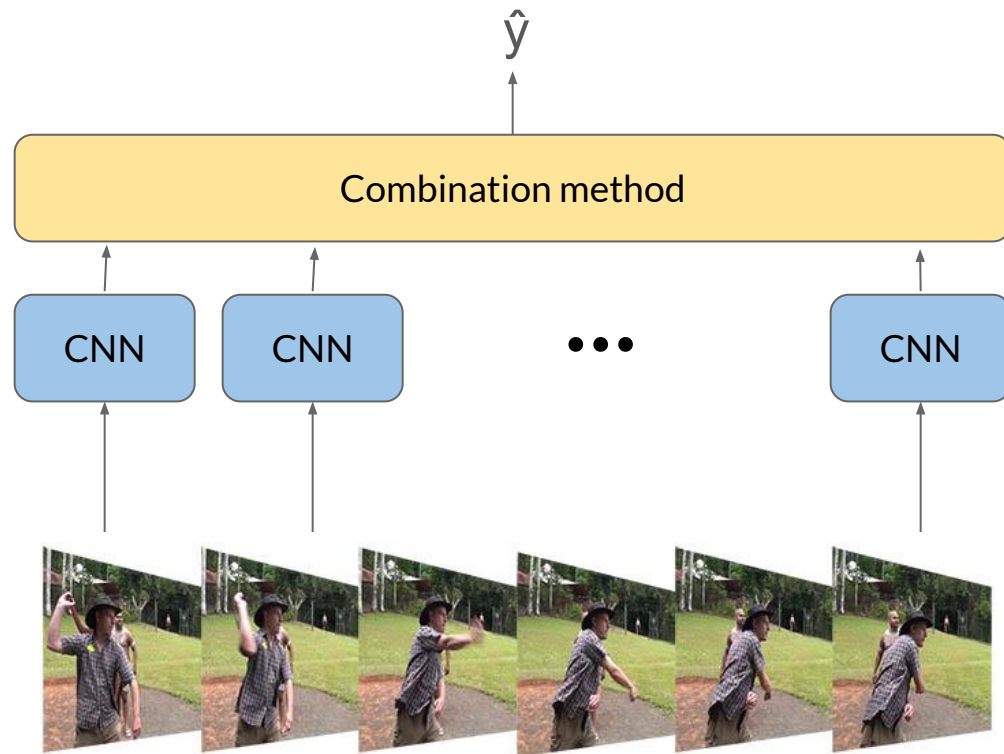
1. Single frame models
2. CNN + RNN
3. 3D convolutions
4. RGB + Optical Flow Two-stream CNN

Deep Video Architectures

Basic deep architectures for video:

1. **Single frame models**
2. CNN + RNN
3. 3D convolutions
4. Two-stream CNN

Single frame models



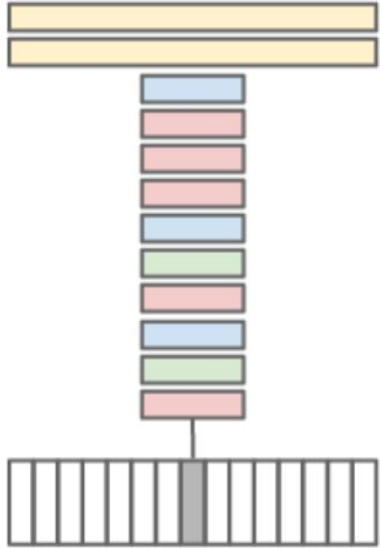
Combination is commonly implemented as a small NN on top of a temporal pooling operation (e.g. max, average).

Problem: pooling is not aware of the temporal order!

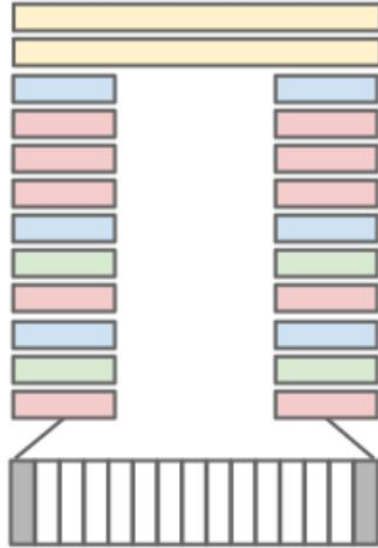
Single frame models



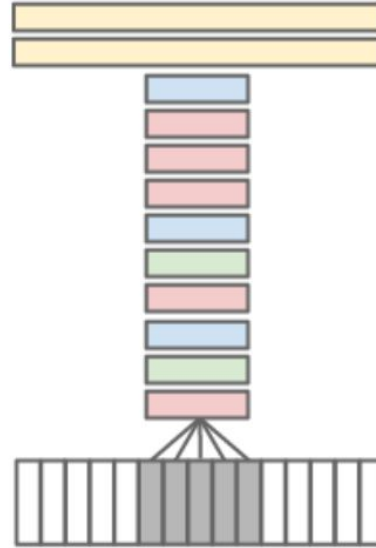
Single Frame



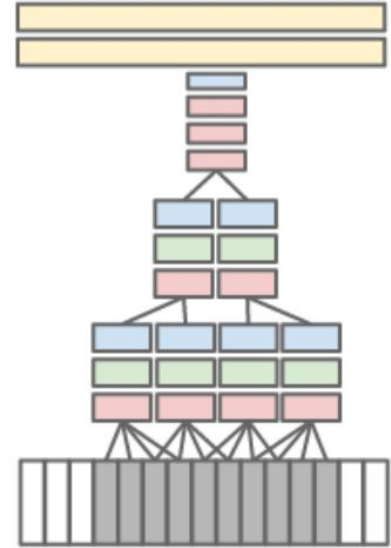
Late Fusion



Early Fusion



Slow Fusion

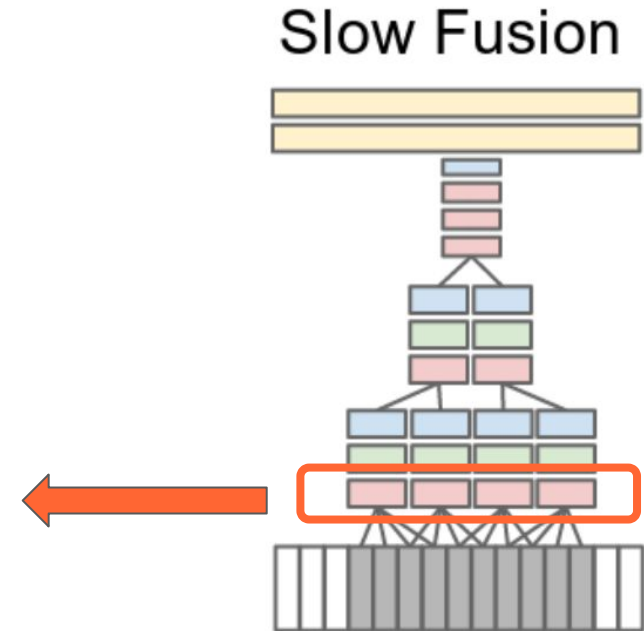
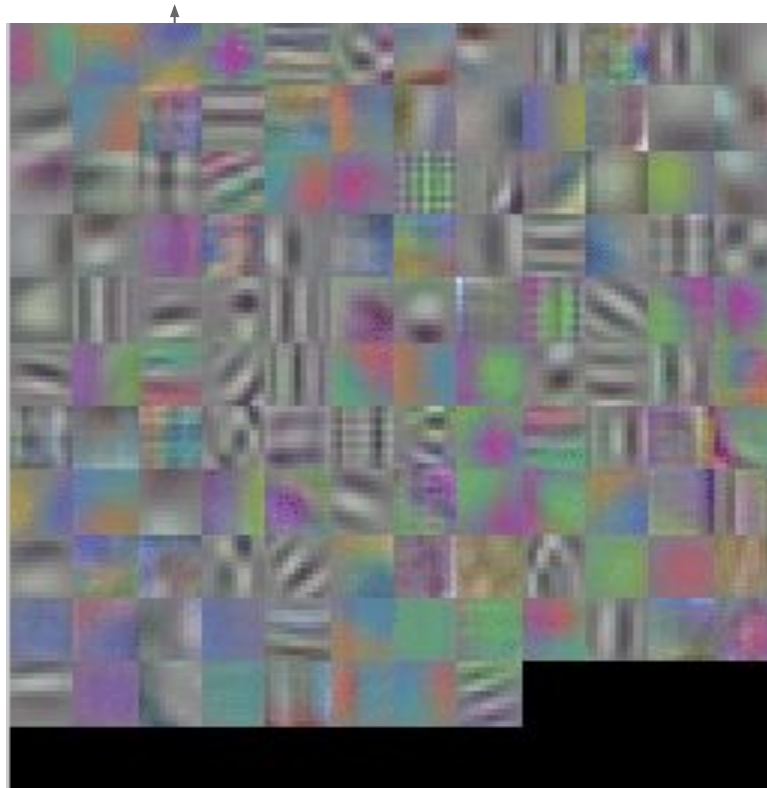


Single frame models



Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	41.9	60.9	80.2
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

Single frame models



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

Deep Video Architectures

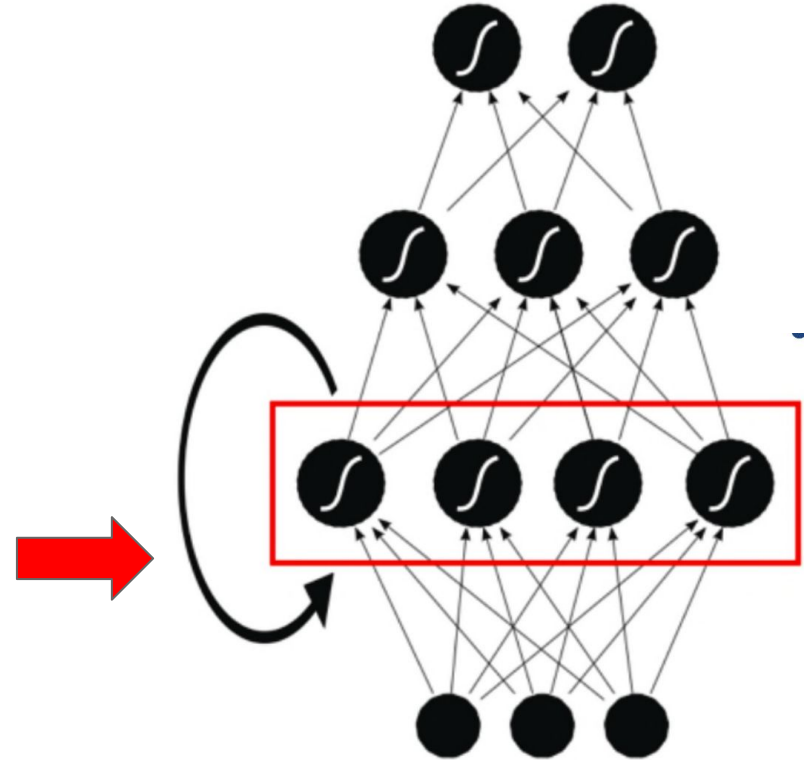
Basic deep architectures for video:

1. Single frame models
2. **CNN + RNN**
3. 3D convolutions
4. RGB + Optical Flow Two-stream CNN

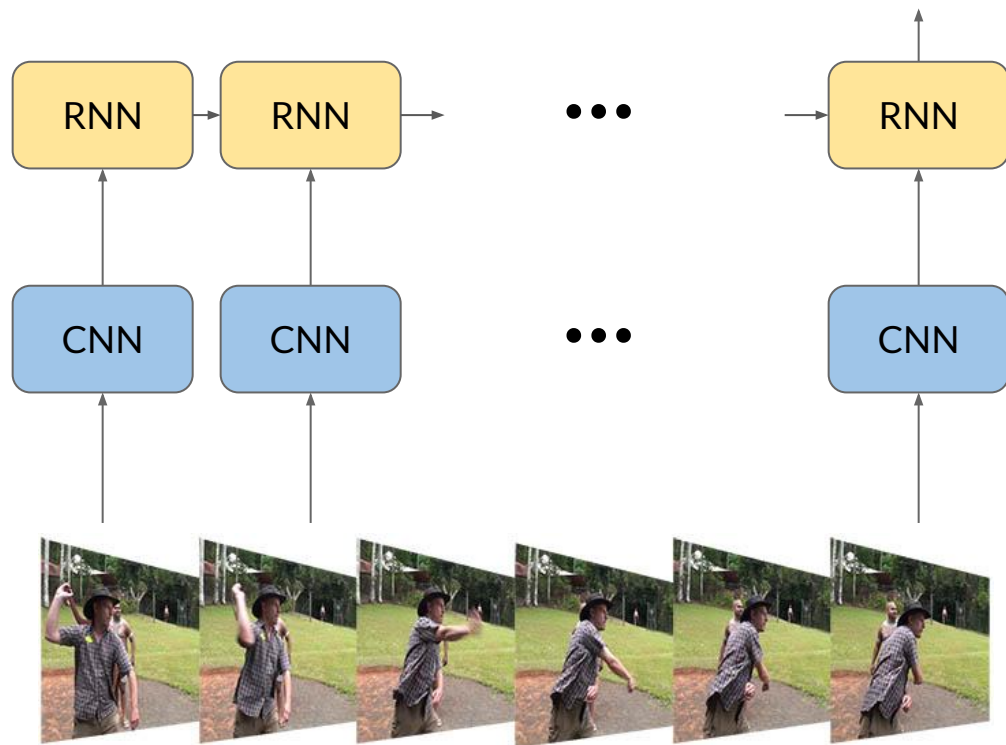
Recurrent Neural Network (RNN)



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



2D CNN + RNN



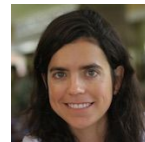
Recurrent Neural Networks are well suited for processing sequences.

Problem: RNNs are sequential and cannot be parallelized.

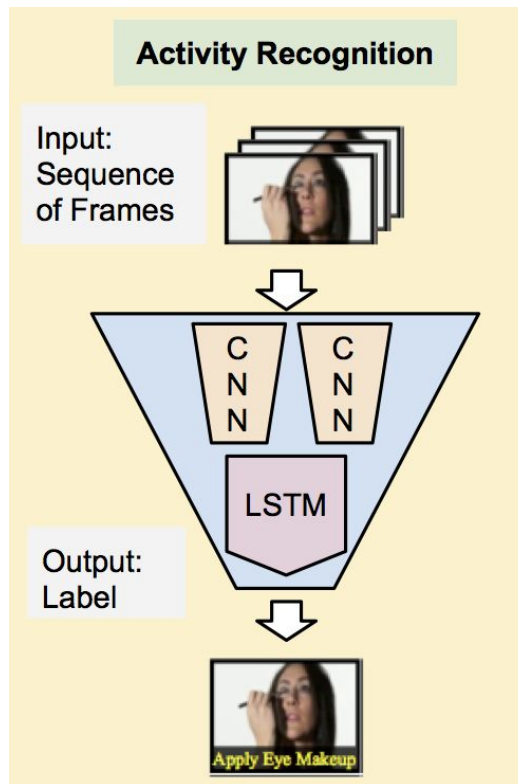
Videolectures on RNNs:

[DLISL 2017](#), ["RNN \(I\)"](#)
["RNN \(II\)"](#)

[DLAI 2018](#), ["RNN"](#)

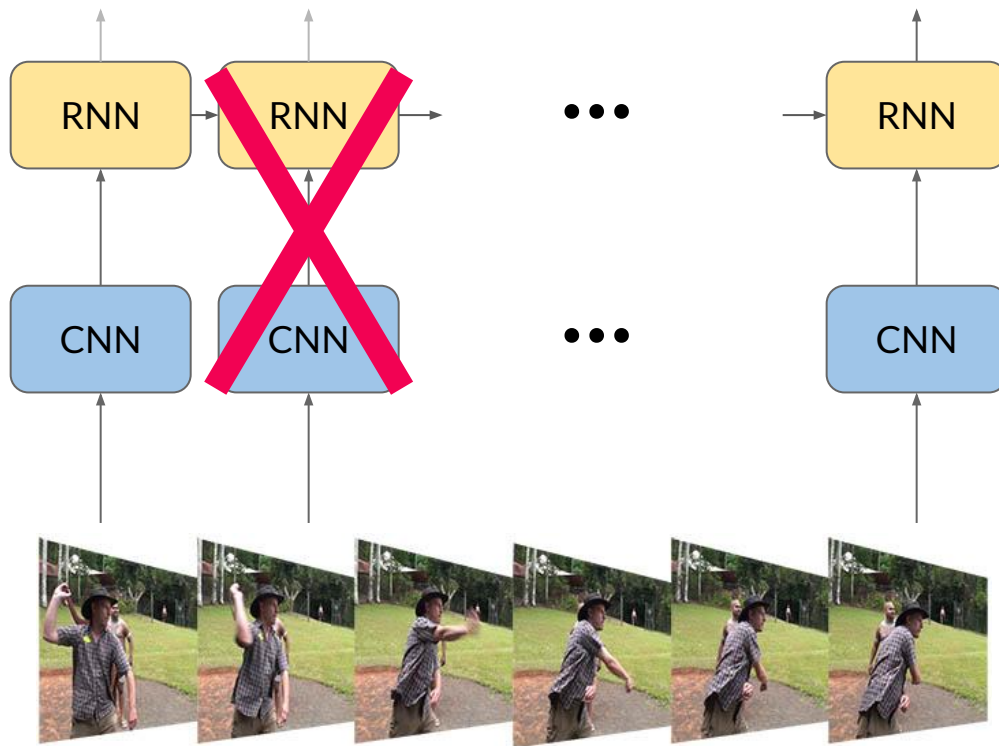


2D CNN + RNN



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](#)

2D CNN + RNN: redundancy



After processing a frame, let the RNN decide how many future frames can be skipped

In skipped frames, simply copy the output and state from the previous time step

There is no ground truth for which frames can be skipped. The RNN **learns** it by itself during training!

2D CNN + RNN: redundancy

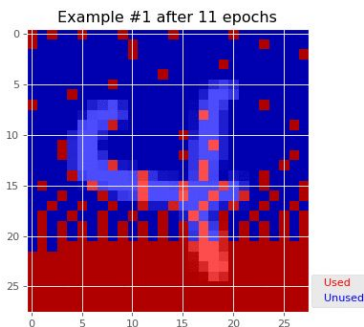


Used
Unused

Victor Campos, Brendan Jou, Xavier Giro-i-Nieto, Jordi Torres, and Shih-Fu Chang. [“Skip RNN: Learning to Skip State Updates in Recurrent Neural Networks”](#), ICLR 2018.

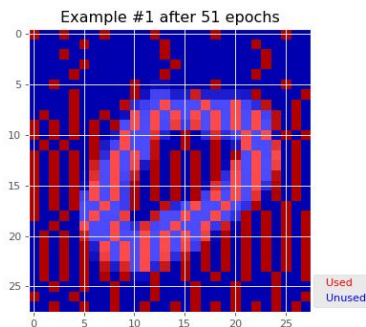
2D CNN + RNN

11 epochs



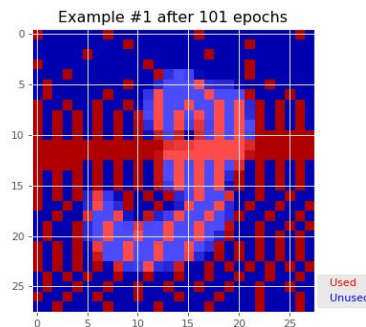
~30% acc

51 epochs



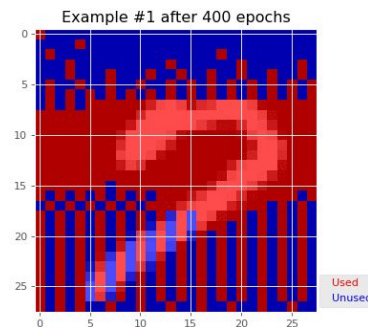
~50% acc

101 epochs



~70% acc

400 epochs



~95% acc

Epochs for Skip LSTM ($\lambda = 10^{-4}$)

Used
Unused

Victor Campos, Brendan Jou, Xavier Giro-i-Nieto, Jordi Torres, and Shih-Fu Chang. [“Skip RNN: Learning to Skip State Updates in Recurrent Neural Networks”](#), ICLR 2018.

Deep Video Architectures

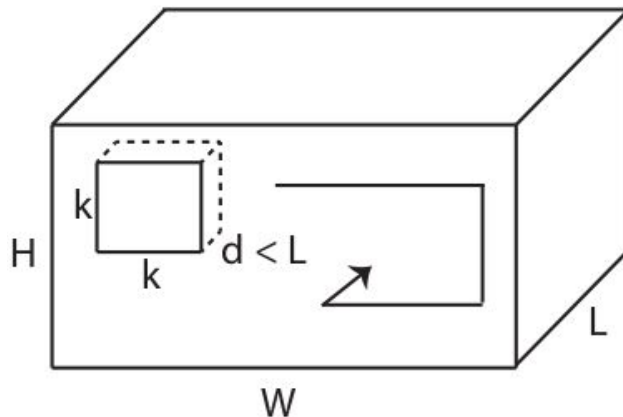
Basic deep architectures for video:

1. Single frame models
2. CNN + RNN
- 3. 3D convolutions**
4. RGB + Optical Flow Two-stream CNN

3D CNN (C3D)

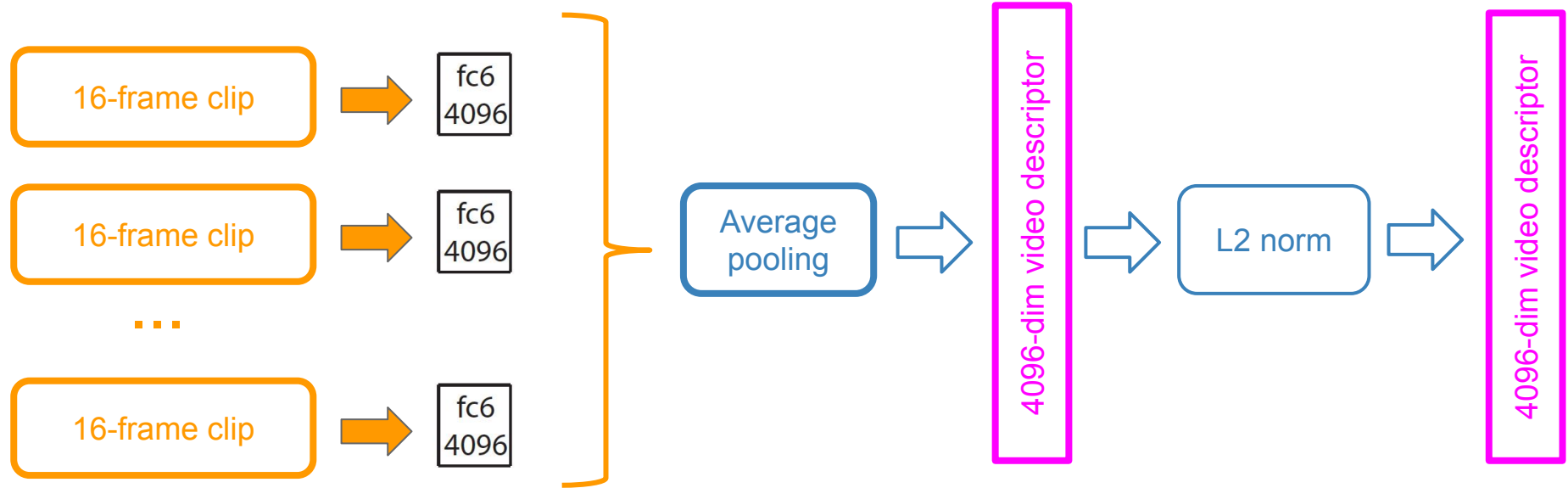
We can add an extra dimension to standard 2D CNN filters:

- A video may be represented by D feature maps of size $H \times W \times L$.
- A 3D conv filter may be of size $k \times k \times d$



3D CNN (C3D) + temporal pooling

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.



3D CNN

Limitation:

- How can we use pre-trained 2D networks to initialize C3D training ?



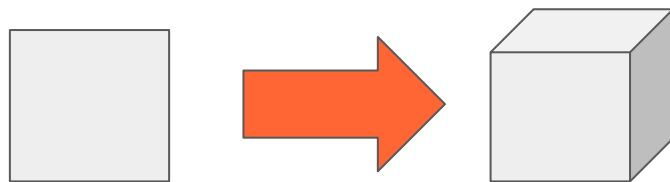
Re-use pre-trained 2D CNNs for 3D CNNs

We can add an extra dimension to standard CNNs:

- An image is a $H \times W \times D$ tensor: $M \times N \times D'$ conv filters
- A video is a $T \times H \times W \times D$ tensor: $K \times M \times N \times D'$ conv filters

We can convert an $M \times N \times D'$ conv filter into a $K \times M \times N \times D'$ filter by replicating it K times in the time axis and scaling its values by $1/K$.

- This allows to leverage networks pre-trained on ImageNet and alleviate the computational burden associated to training from scratch

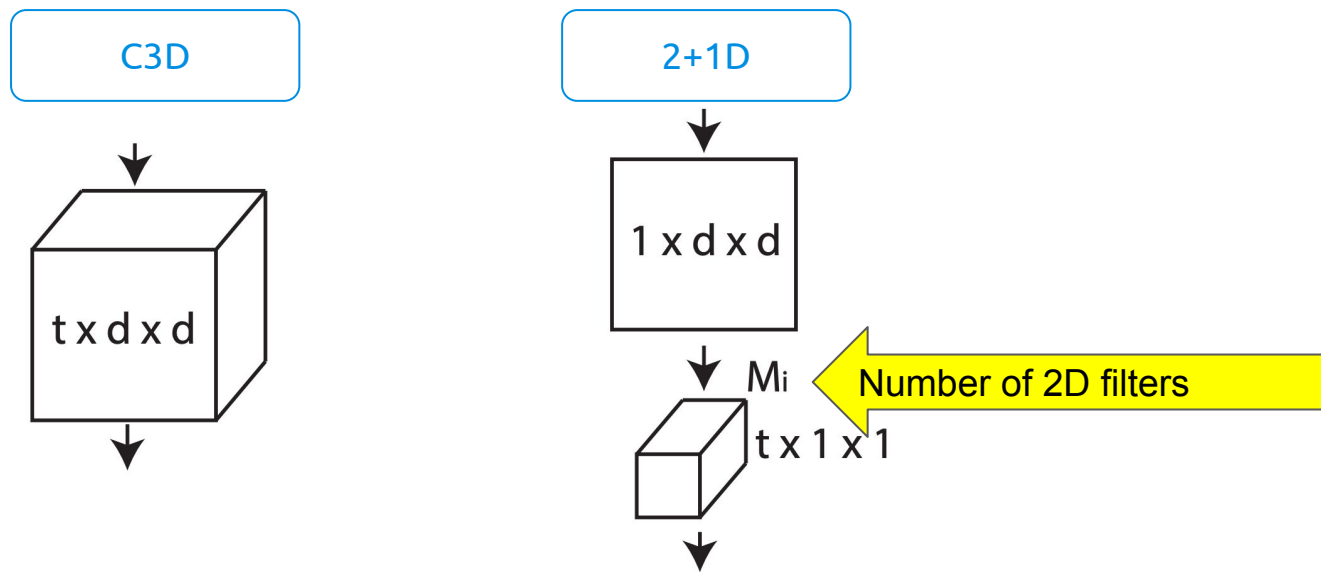


3D CNN + Residual (Res3D)

Dataset	Sports1M acc.(%)	UCF101 acc.(%)	HMDB51 acc.(%)	THU14 mAP(%)	ASLAN acc(%)
C3D	61.1	82.3	51.6	19.0	78.3
Res3D	65.6	85.8	54.9	22.5	78.8
Δ	4.5	3.5	3.3	3.5	0.5

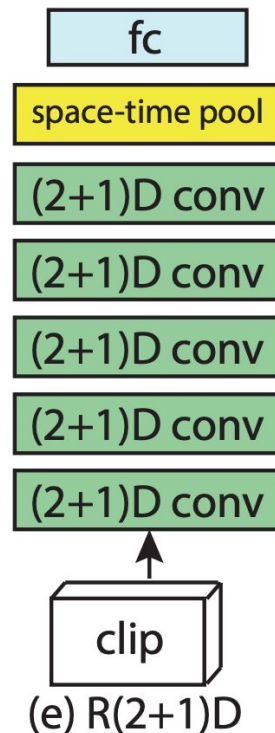
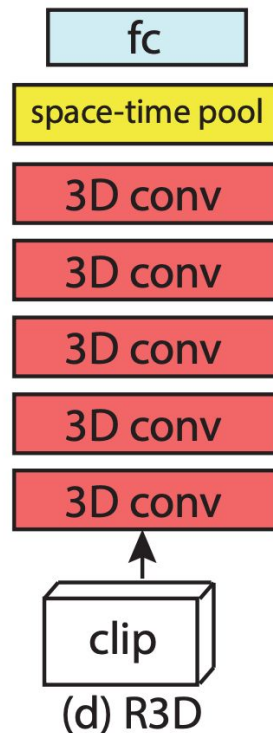
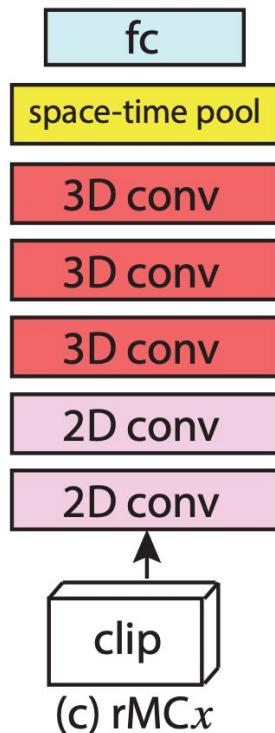
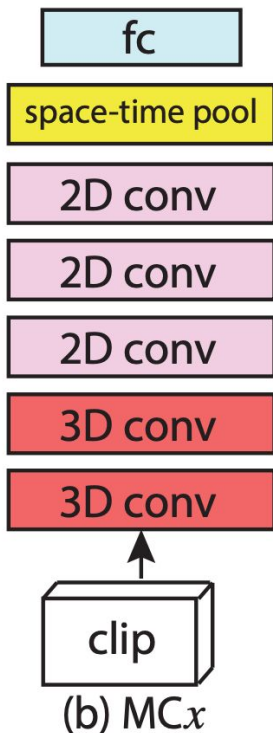
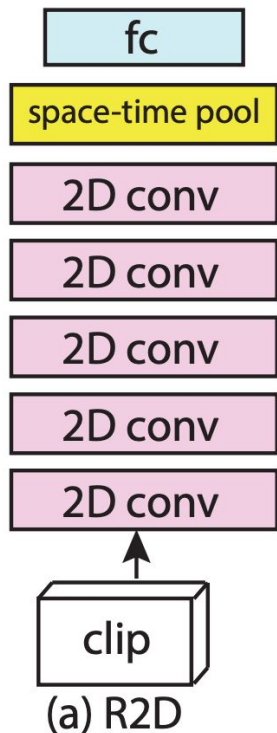
2+1D CNN

A (2+1)D convolutional block splits the computation into a spatial 2D convolution followed by a temporal 1D convolution. In the case of an input of a single feature channel:



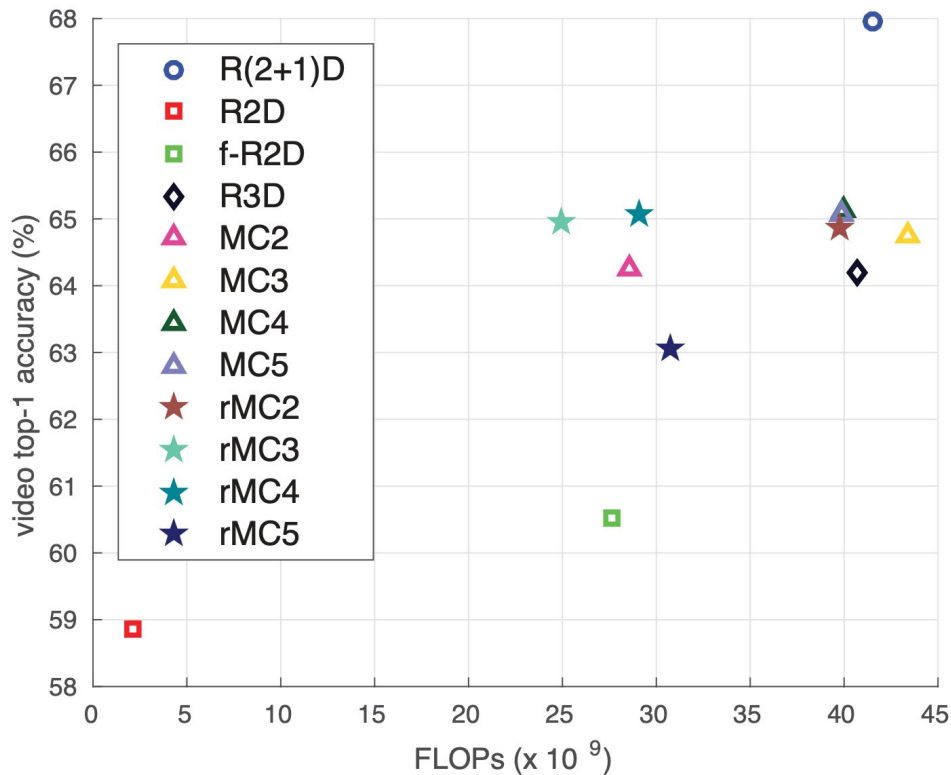
#R(2+1)D Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. ["A closer look at spatiotemporal convolutions for action recognition."](#) CVPR 2018. [\[code\]](#)

2+1D CNN + Residual



#R(2+1)D Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. ["A closer look at spatiotemporal convolutions for action recognition."](#) CVPR 2018. [\[code\]](#)

2+1D CNN + Residual

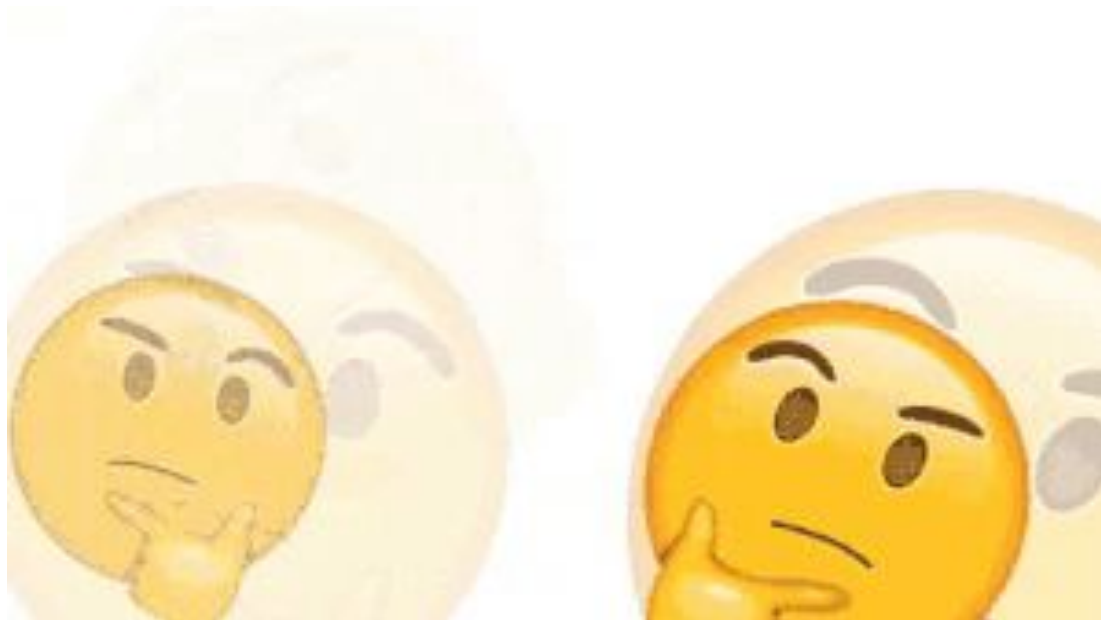


#R(2+1)D Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. ["A closer look at spatiotemporal convolutions for action recognition."](#) CVPR 2018. [\[code\]](#)

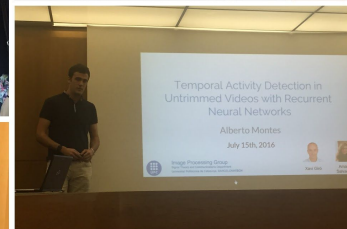
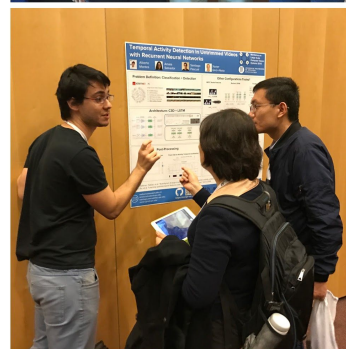
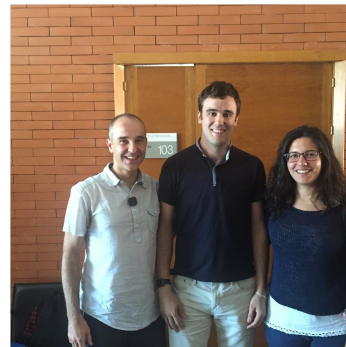
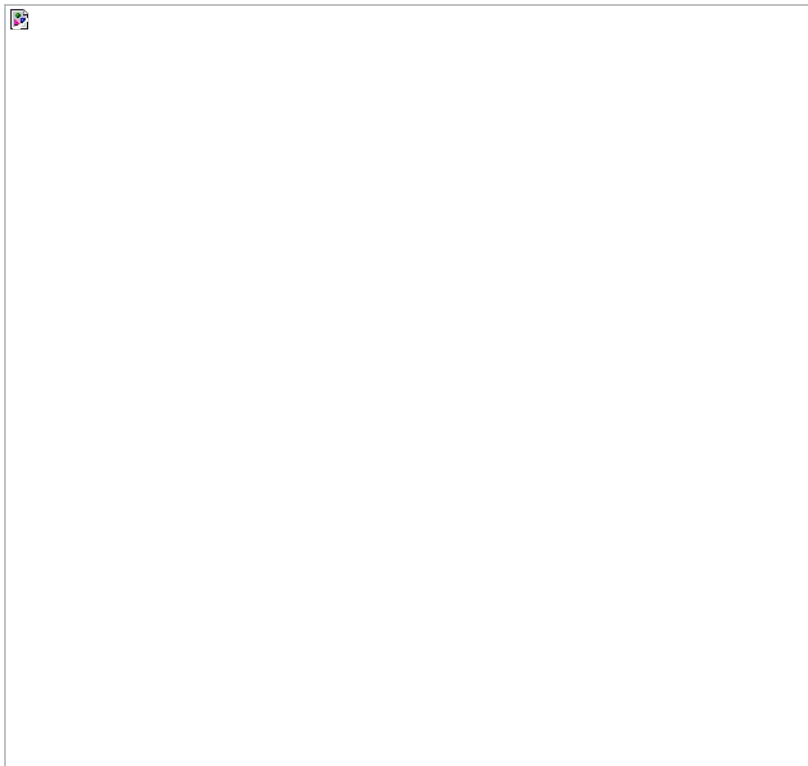
3D CNN

Limitations:

- How can we handle longer videos?
- How can we capture longer temporal dependencies?



3D CNN + RNN



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)*

Deep Video Architectures

Basic deep architectures for video:

1. Single frame models
2. CNN + RNN
3. 3D convolutions
4. **RGB + Optical Flow Two-stream CNN**

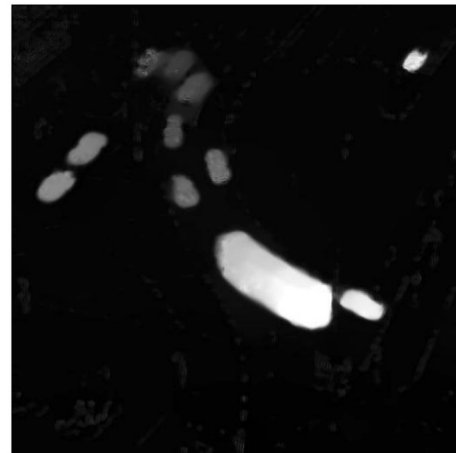


Ilg, Eddy, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. "[FlowNet 2.0: Evolution of optical flow estimation with deep networks.](#)" CVPR 2017. [code]

Optical flow

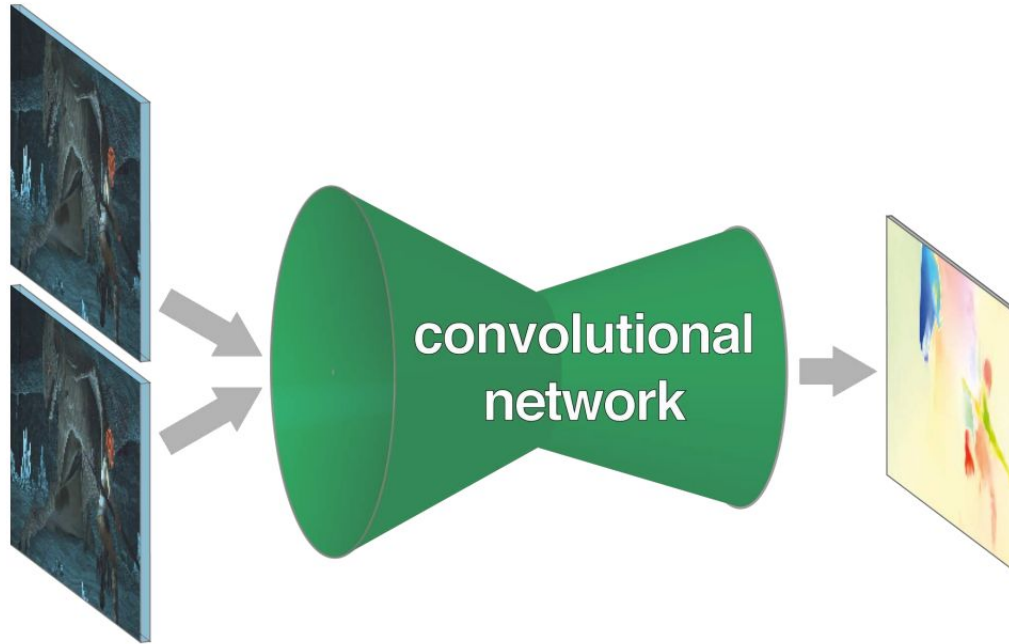
Popular implementations for optical flow:

- [PyFlow](#)
- [FlowNet2](#)
- [Improved Dense Trajectories - IDT](#)
- [Lucas-Kanade](#) (OpenCV)
- [Farneback 2003](#) (OpenCV)
- [Middlebury](#)
- [Horn and schunk](#)
- [Tikhonov regularized and vectorized](#)
- [DeepFlow](#) (2013)
- (...)



Optical flow

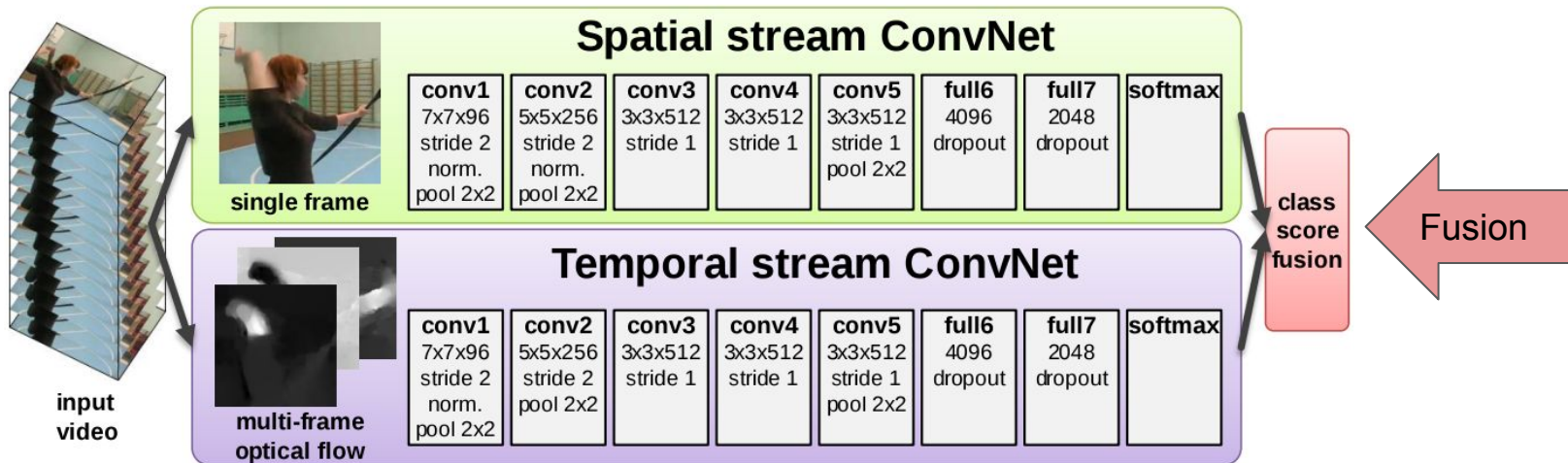
Deep Neural Networks have actually also been trained to predict optical flow:



Two-streams 2D CNNs

Problem: Single frame models do not take into account motion in videos.

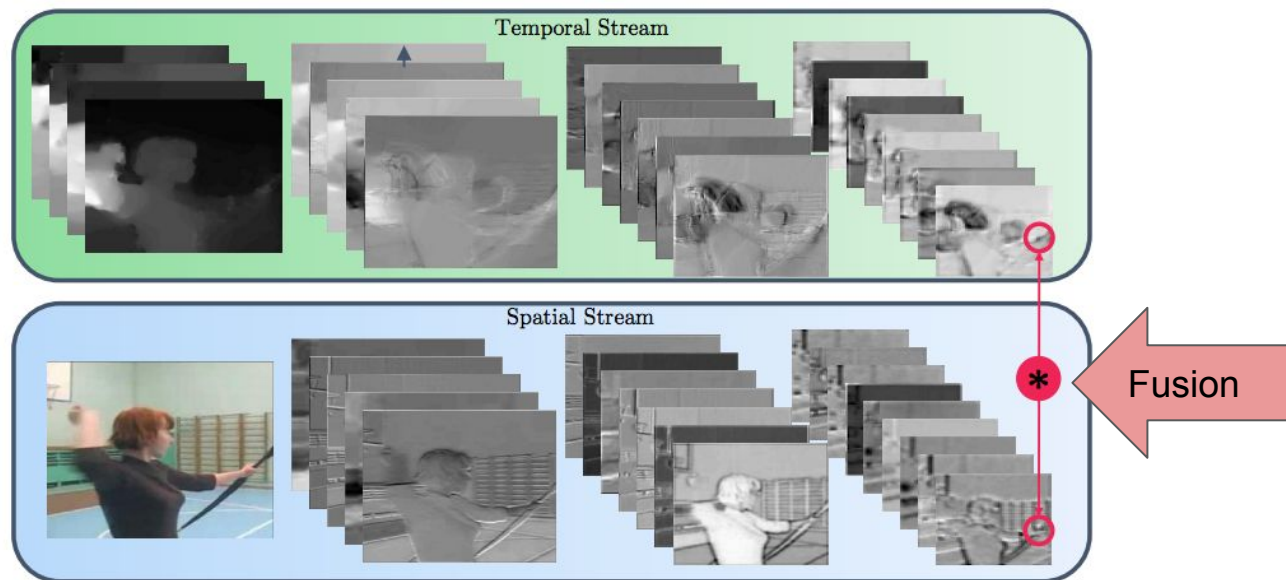
Solution: extract **optical flow** for a stack of frames and use it as an input to a CNN.



Two-streams 2D CNNs

Problem: Single frame models do not take into account motion in videos.

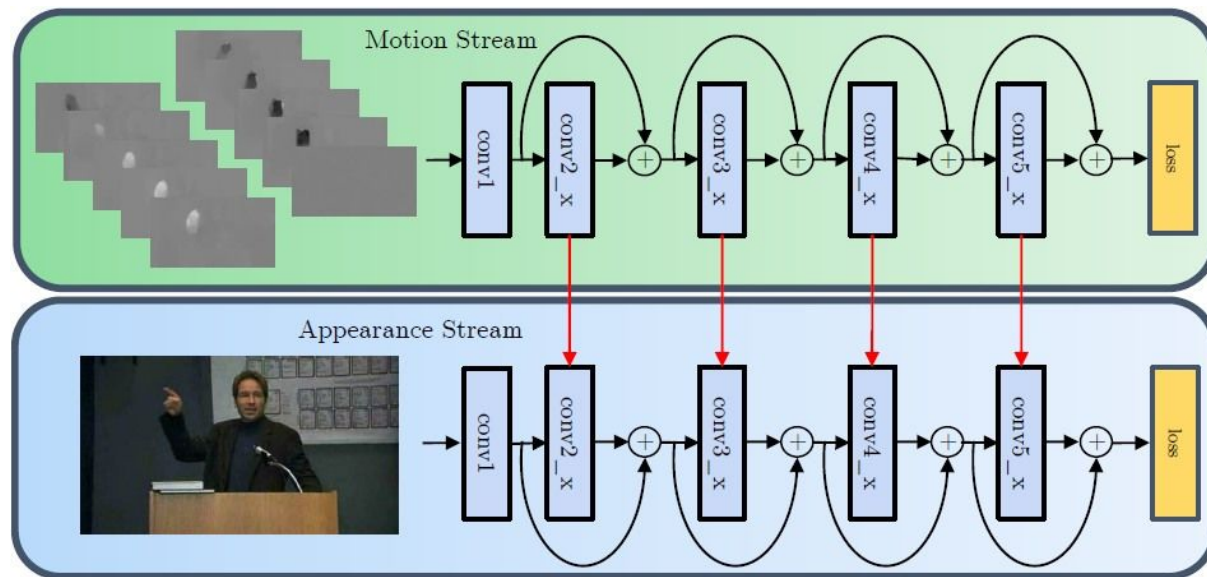
Solution: extract optical flow for a stack of frames and use it as an input to a CNN.



Two-streams 2D CNNs

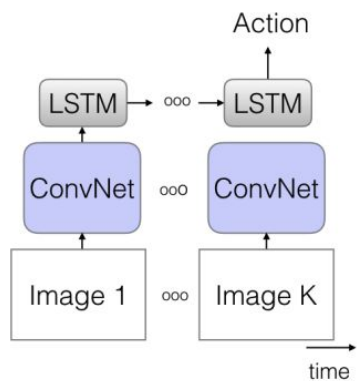
Problem: Single frame models do not take into account motion in videos.

Solution: extract optical flow for a stack of frames and use it as an input to a CNN.

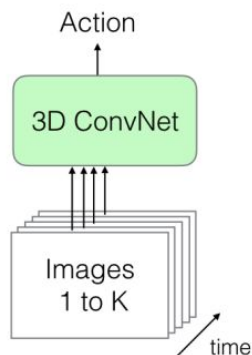


Two-streams 3D CNNs

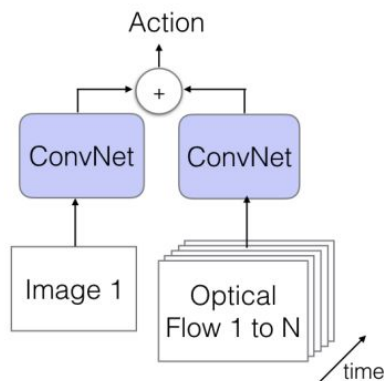
a) LSTM



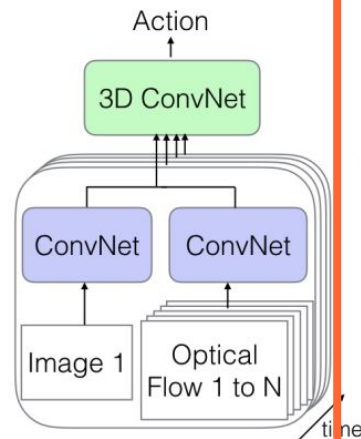
b) 3D-ConvNet



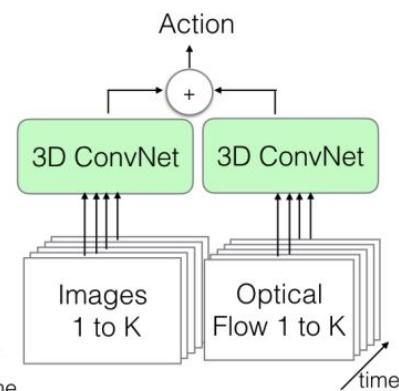
c) Two-Stream



d) 3D-Fused Two-Stream

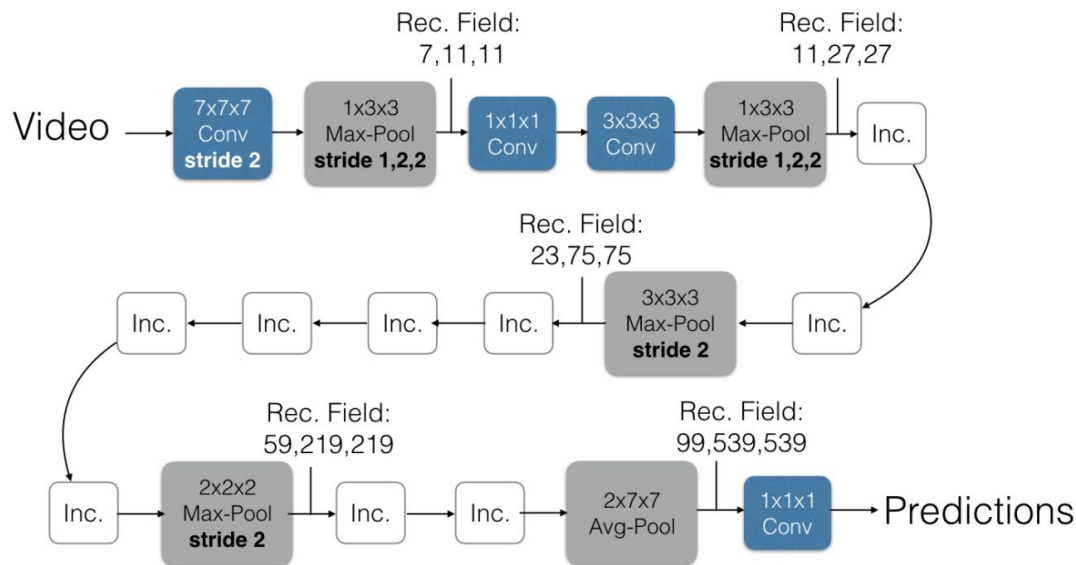


e) Two-Stream 3D-ConvNet

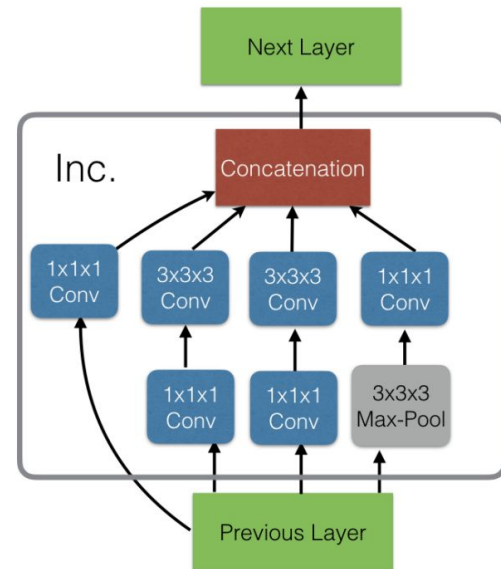


Two-streams 3D CNNs

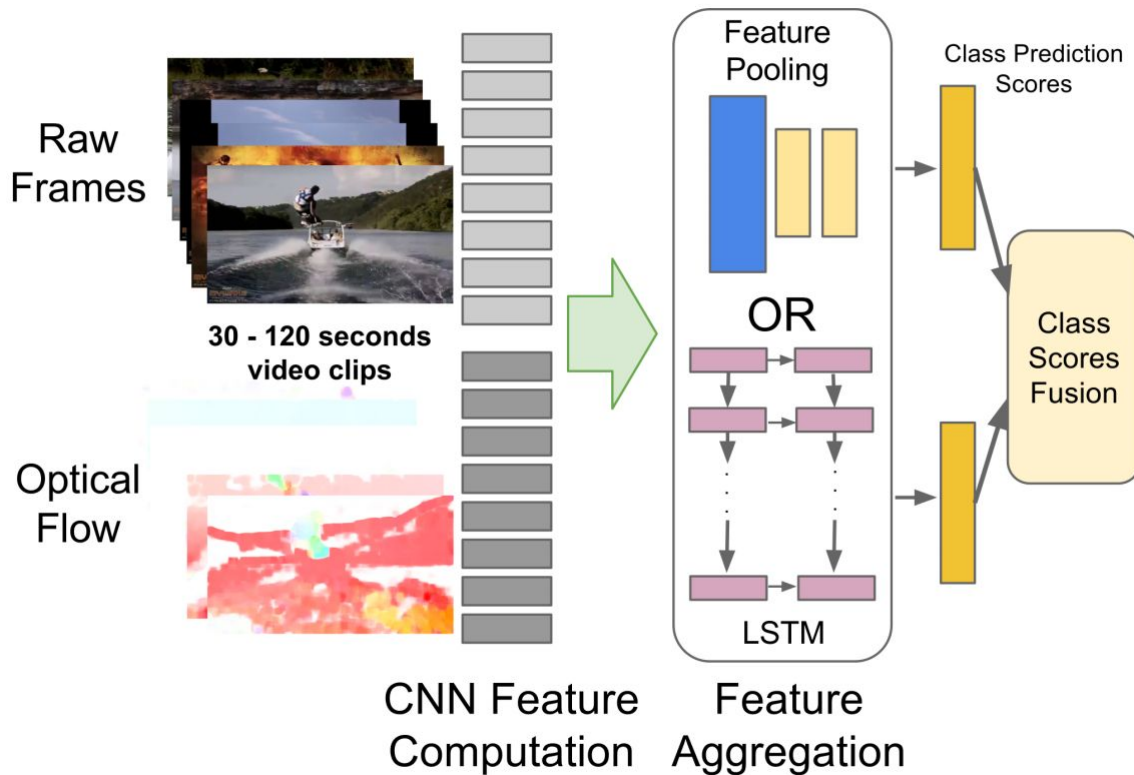
Inflated Inception-V1



Inception Module (Inc.)



Two-streams 2D CNNs + RNN



Outline

1. Architectures
- 2. Tips and tricks**

Large-scale datasets

- The reference dataset for image classification, ImageNet, has ~1.3M images
 - Training a state of the art CNN can take up to 2 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



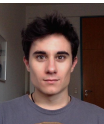
Memory issues

- Current GPUs can fit batches of 32~64 images when training state of the art CNNs
 - This means 32~64 video frames at once
- Memory footprint can be reduced in different ways if a pre-trained CNN model is used
 - Freezing some of the lower layers, reducing the memory impact of backprop
 - Extracting frame-level features and training a model on top of it (e.g. RNN on top of CNN features). This is equivalent to freezing the whole architecture, but the CNN part needs to be computed only once.



I/O bottleneck

- In practice, applying deep learning to video analysis requires from multi-GPU or distributed settings
- In such settings it is very important to avoid *starving* the GPUs or we will not obtain any speedup
 - The next batch needs to be loaded and preprocessed to keep the GPU as busy as possible
 - Using asynchronous data loading pipelines is a key factor
 - Loading individual files is slow due to the introduced overhead, so using other formats such as TFRecord/HDF5/LMDB is highly recommended





Qure.ai Blog

Revolutionizing healthcare with deep learning



Deep Learning for Videos: A 2018 Guide to Action Recognition

Rohit Ghosh | June 11, 2018

Medical images like MRIs, CTs (3D images) are very similar to videos - both of them encode 2D spatial information over a 3rd dimension. Much like diagnosing abnormalities from 3D images, action recognition from videos would require capturing context from entire video rather than just capturing information from each frame.



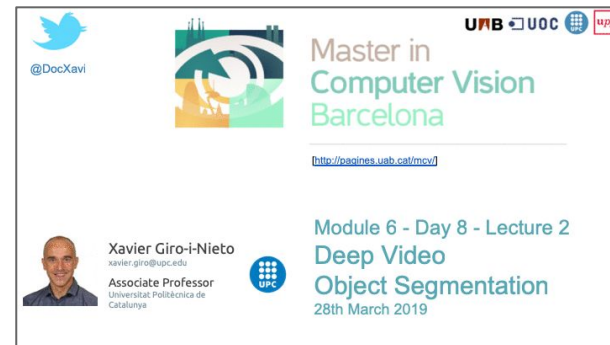
Fig 1: Left: Example Head CT scan. Right: Example video from an action recognition dataset. Z dimension in the CT volume is analogous to time dimension in the video.

Choose your next lecture [\[index\]](#)



This slide thumbnail features the Twitter handle @DocXavi, the Master in Computer Vision Barcelona logo, and logos for UMB, UOC, and UPF. It includes a portrait of Xavier Giro-i-Nieto and the text: "Module 6 - Day 9 - Lecture 1", "Deep Video", "Object Tracking", and "4th April 2019". The footer identifies him as an Associate Professor at the Department of Signal Theory and Communications, Image Processing Group, UPC.

Object tracking
[\[slides\]](#) [\[video\]](#)



This slide thumbnail features the Twitter handle @DocXavi, the Master in Computer Vision Barcelona logo, and logos for UMB, UOC, and UPF. It includes a portrait of Xavier Giro-i-Nieto and the text: "Module 6 - Day 8 - Lecture 2", "Deep Video", "Object Segmentation", and "28th March 2019". The footer identifies him as an Associate Professor at the Universitat Politècnica de Catalunya.

Video Object segmentation
[\[slides\]](#) [\[video\]](#)



This slide thumbnail features the Twitter handle @DocXavi, the Master in Computer Vision Barcelona logo, and logos for UMB, UOC, and UPF. It includes a portrait of Xavier Giro-i-Nieto and the text: "Module 6 - Day 8 - Lecture 1", "Self-supervised Learning", "from Video Sequences", and "28th March 2019". The footer identifies him as an Associate Professor at the Universitat Politècnica de Catalunya.

Self-supervised Learning from
Video Sequences [\[slides\]](#) [\[video\]](#)



This slide thumbnail features the Twitter handle @DocXavi, the Master in Computer Vision Barcelona logo, and logos for UMB, UOC, and UPF. It includes a portrait of Xavier Giro-i-Nieto and the text: "Module 6 - Day 9 - Lecture 2", "Self-supervised", "Audiovisual Learning", and "4th April 2019". The footer identifies him as an Associate Professor at the Universitat Politècnica de Catalunya.

Self-supervised Audiovisual
Learning [\[slides\]](#) [\[video\]](#)

Questions ?

Undergradese

What undergrads ask vs. what they're REALLY asking

"Is it going to be an open book exam?"

Translation: "I don't have to actually memorize anything, do I?"

"Hmm, what do you mean by that?"

Translation: "What's the answer so we can all go home."

"Are you going to have office hours today?"

Translation: "Can I do my homework in your office?"

"Can i get an extension?"

Translation: "Can you re-arrange your life around mine?"

"Is this going to be on the test?"

Translation: "Tell us what's going to be on the test."

"Is grading going to be curved?"

Translation: "Can I do a mediocre job and still get an A?"

JORGE CHAM © 2008

