@DocXavi

Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giró-i-Nieto

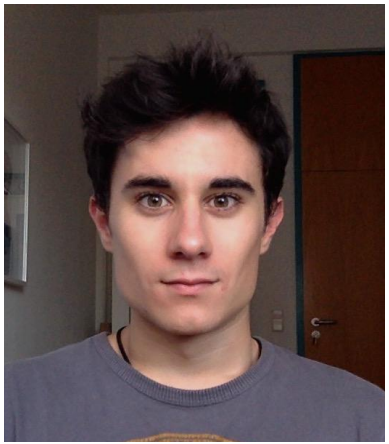
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

Module 6

Deep Learning for Video: Action Recognition

22nd March 2018

Acknowledgements



Víctor Campos

Amaia Salvador

Alberto Montes

Santiago Pascual

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Densely linked slides

Outline

1. **Architectures**
2. Datasets
3. 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

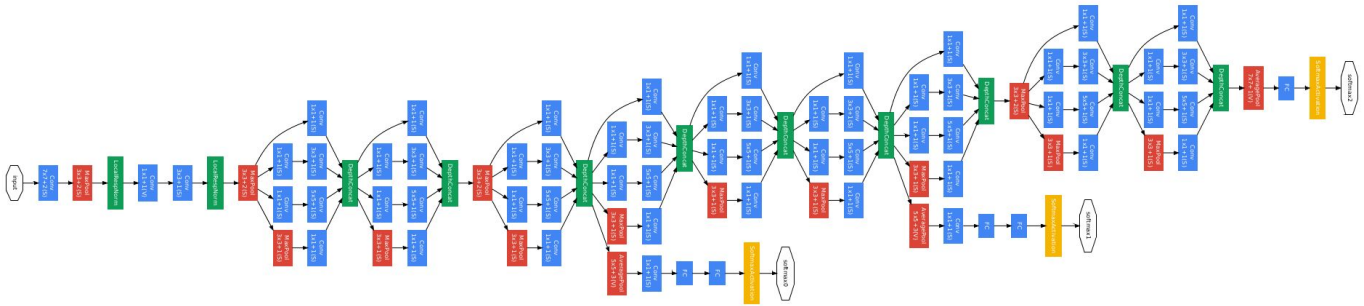
Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. Large-scale video classification with convolutional neural networks. CVPR 2014.

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)

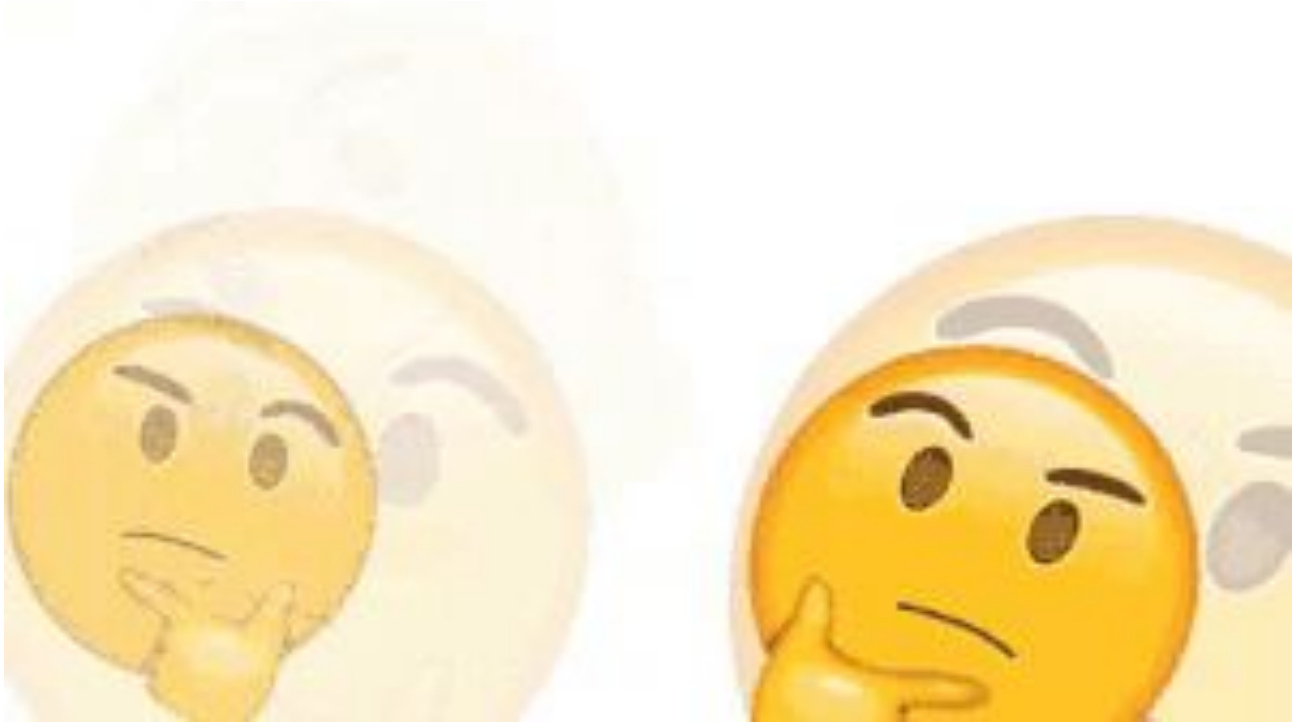
How do we work with images?

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



How do we work with videos ?

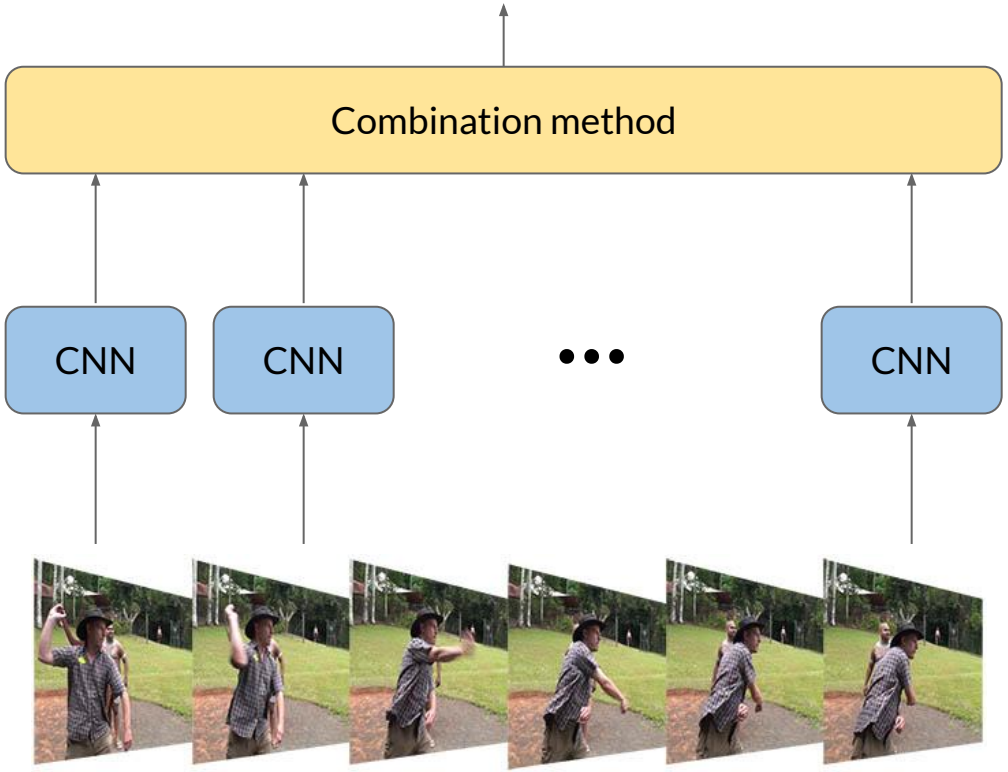
- How can we extend CNNs to image sequences?



CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling

Single frame models



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

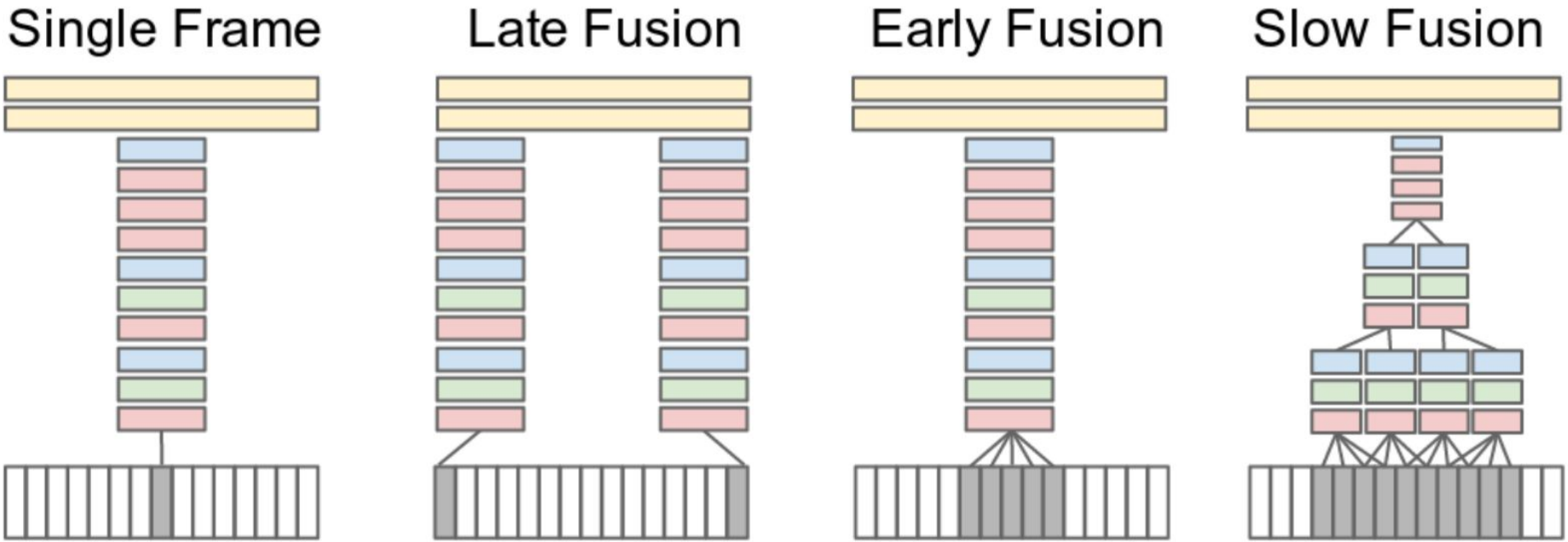
Problem: pooling is not aware of the temporal order!

Yue-Hei Ng, Joe, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. "Beyond short snippets: Deep networks for video classification." CVPR 2015

CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN

Multiple Frames



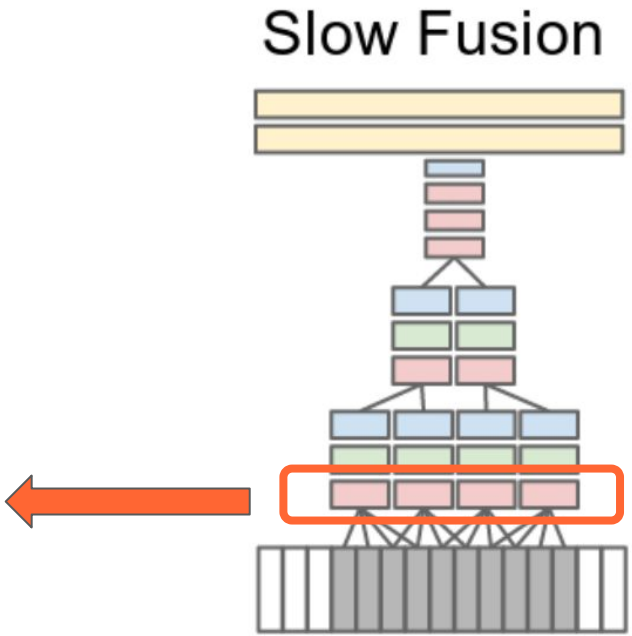
Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. . Large-scale video classification with convolutional neural networks. CVPR 2014

Multiple Frames

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

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

Multiple Frames



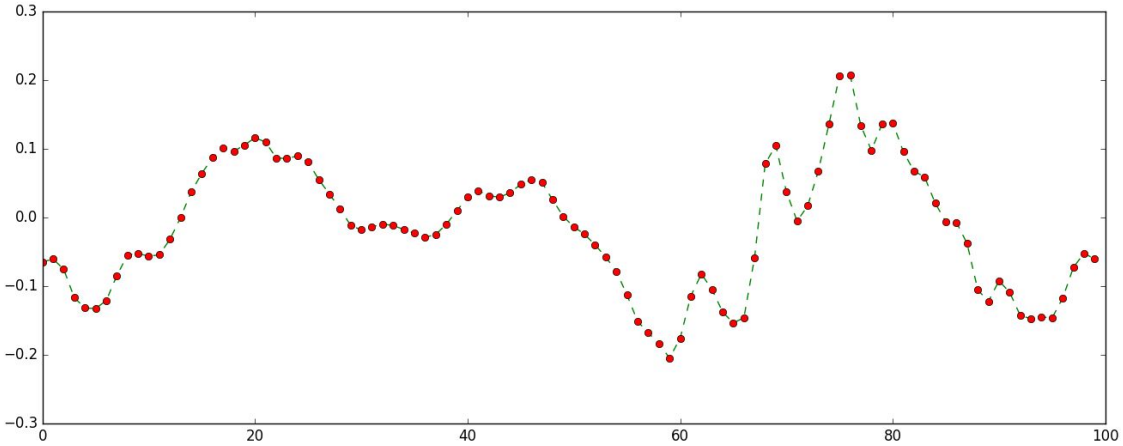
Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. . Large-scale video classification with convolutional neural networks. CVPR 2014

Limitation of Feed Forward NN (as CNNs)

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN

Limitation of Feed Forward NN (as CNNs)

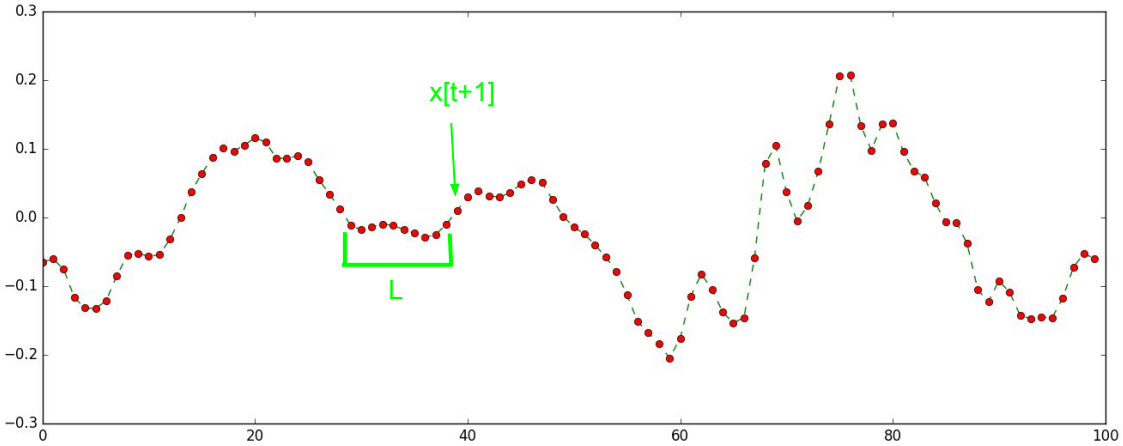
If we have a **sequence** of samples...



predict sample x[t+1] knowing previous values {x[t], x[t-1], x[t-2], …, x[t-τ]}

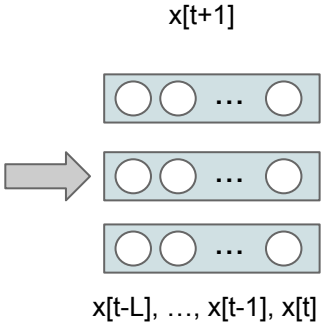
Slide credit: Santi Pascual

Limitation of Feed Forward NN (as CNNs)



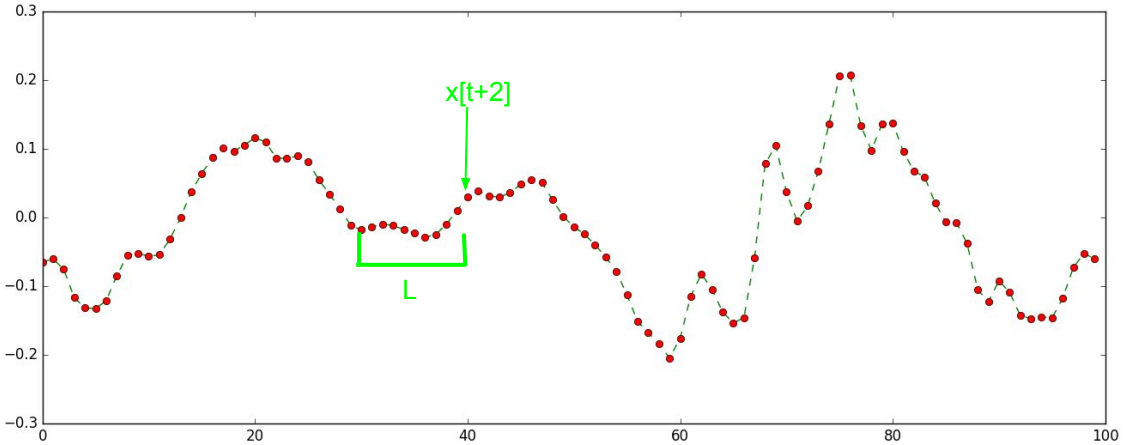
Feed Forward approach:

- static window of size L
- slide the window time-step wise



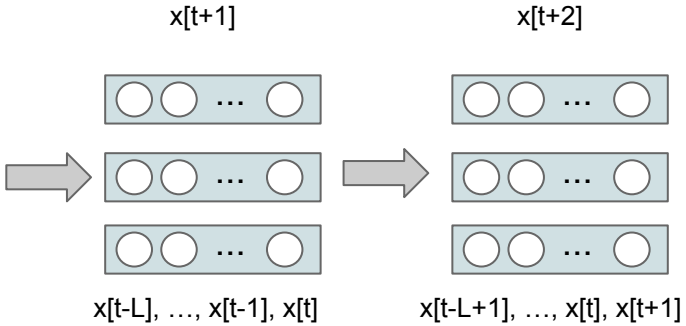
Slide credit: Santi Pascual

Limitation of Feed Forward NN (as CNNs)



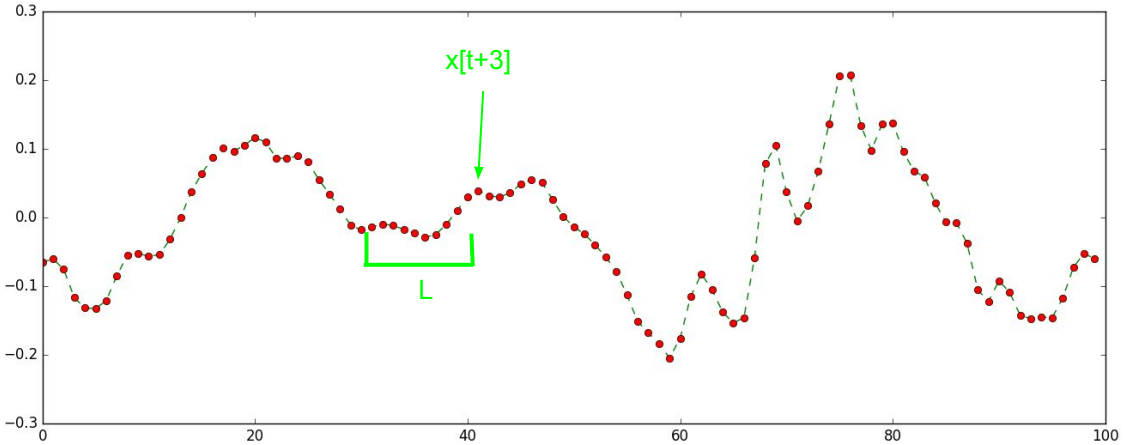
Feed Forward approach:

- static window of size L
- slide the window time-step wise



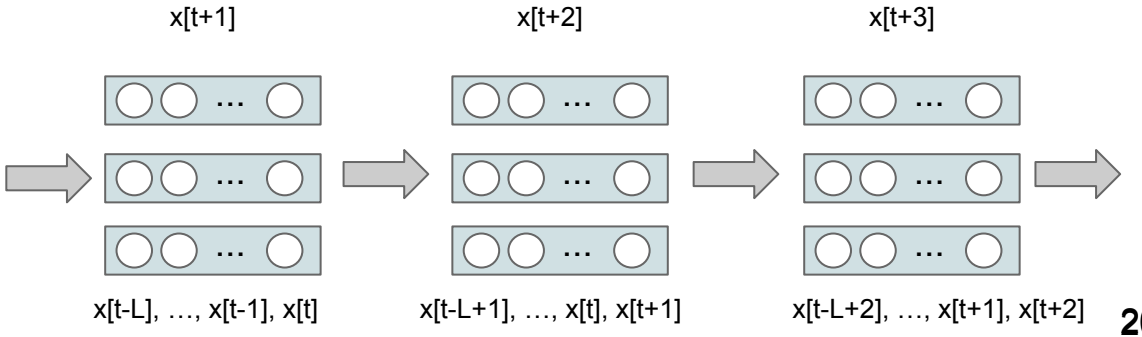
Slide credit: Santi Pascual

Limitation of Feed Forward NN (as CNNs)



Feed Forward approach:

- static window of size L
- slide the window time-step wise

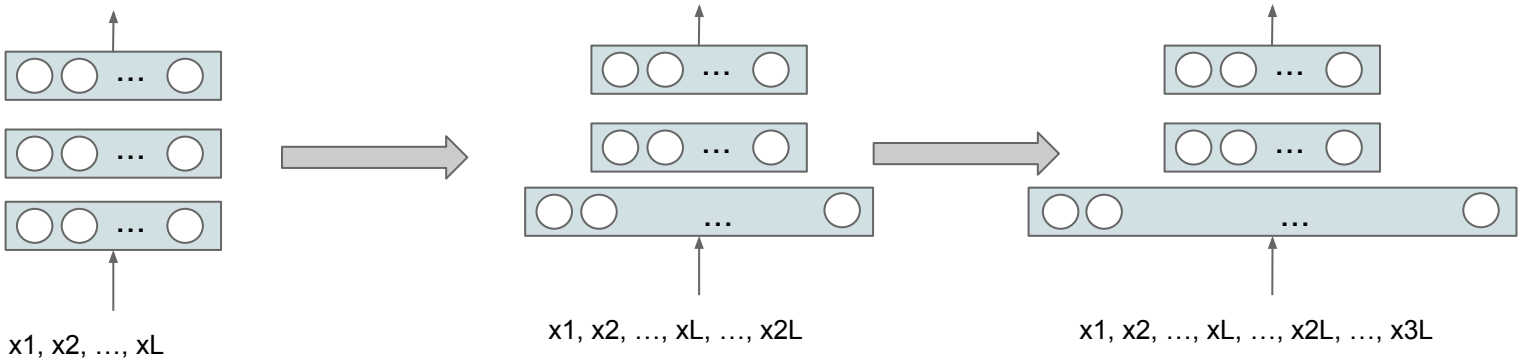


Slide credit: Santi Pascual

Limitation of Feed Forward NN (as CNNs)

Problems for the feed forward + static window approach:

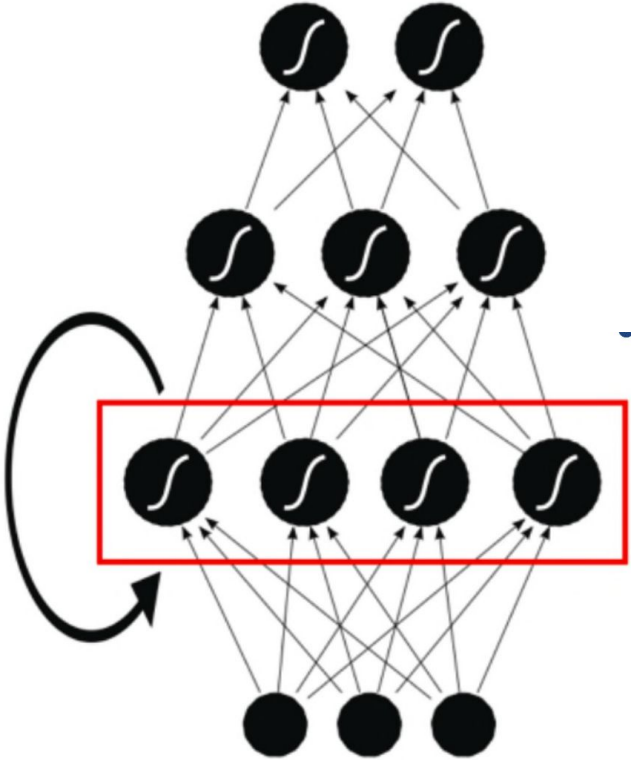
- What's the matter increasing L? → Fast growth of num of parameters!
- Decisions are independent between time-steps!
  - The network doesn't care about what happened at previous time-step, only present window matters → doesn't look good
- Cumbersome padding when there are not enough samples to fill L size
  - Can't work with variable sequence lengths



Slide credit: Santi Pascual

Recurrent Neural Network (RNN)

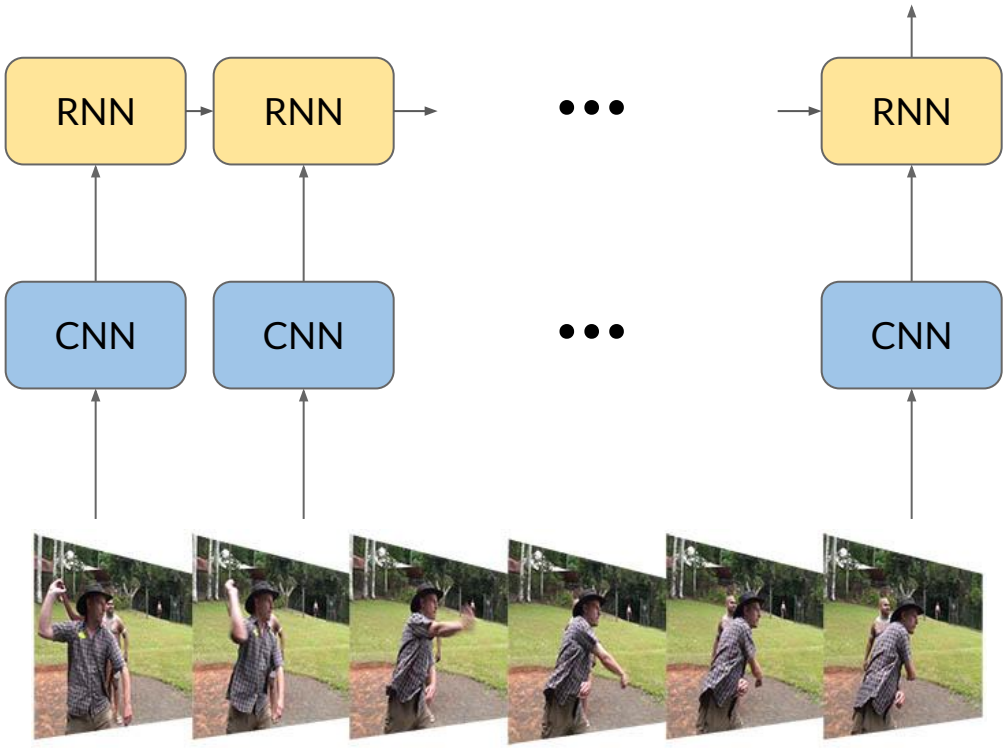
The hidden layers and the output depend from previous states of the hidden layers



CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN

2D CNN + RNN



Recurrent Neural Networks are well suited for processing sequences.

Problem: RNNs are sequential and cannot be parallelized.

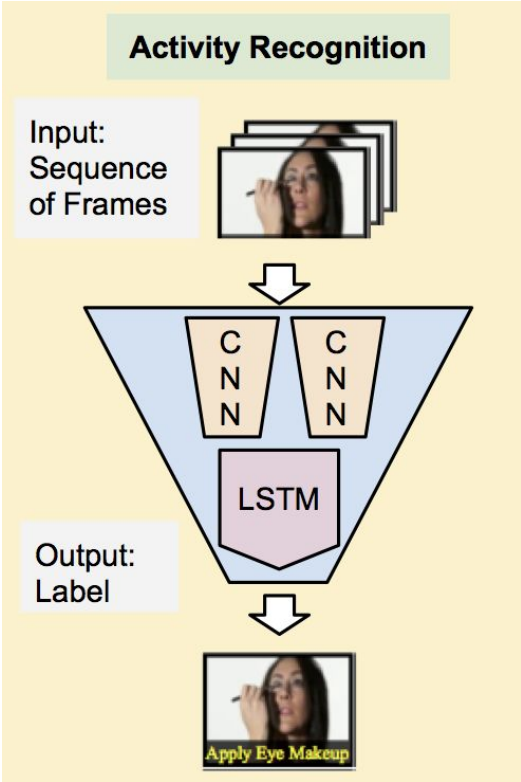
Videolectures on RNNs:

DLSL 2017, "RNN (I)" "RNN (II)"

DLAI 2018, "RNN"

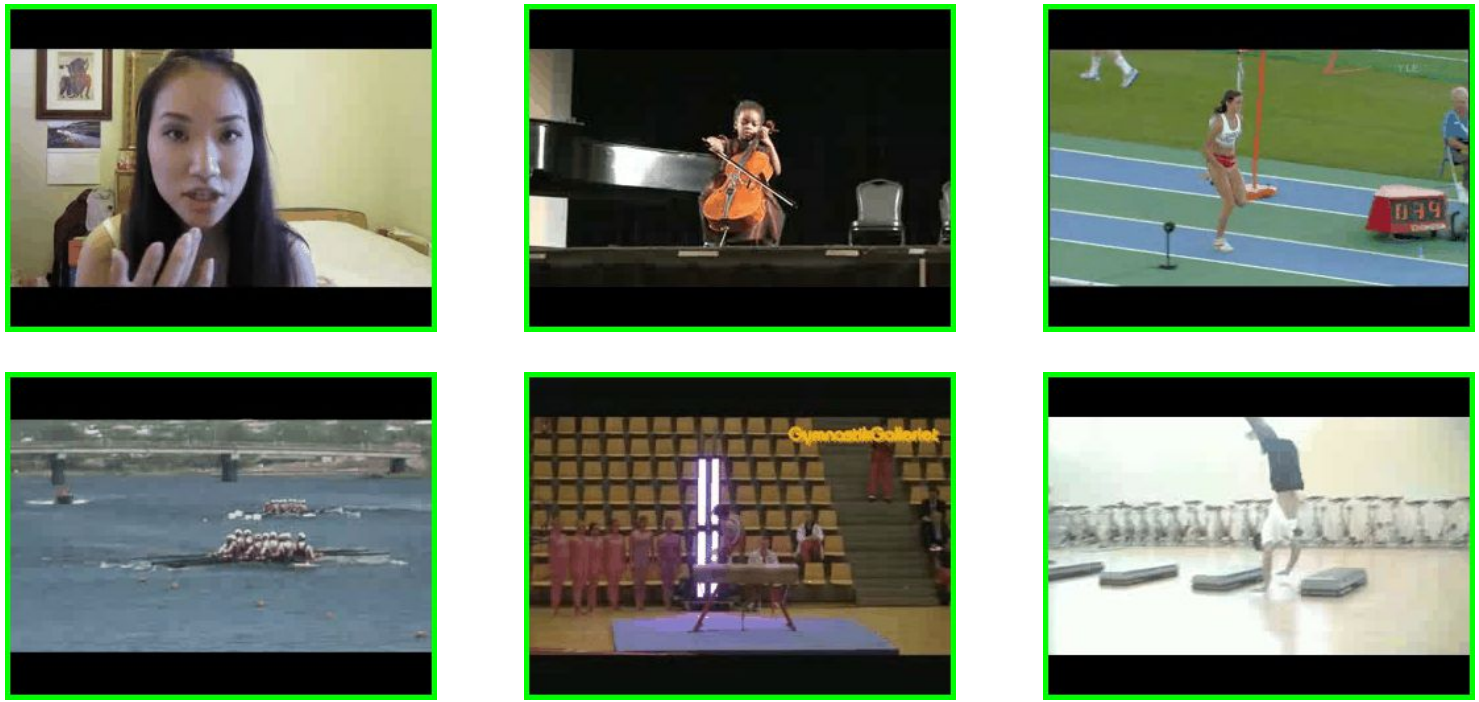
Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

2D CNN + RNN



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



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.

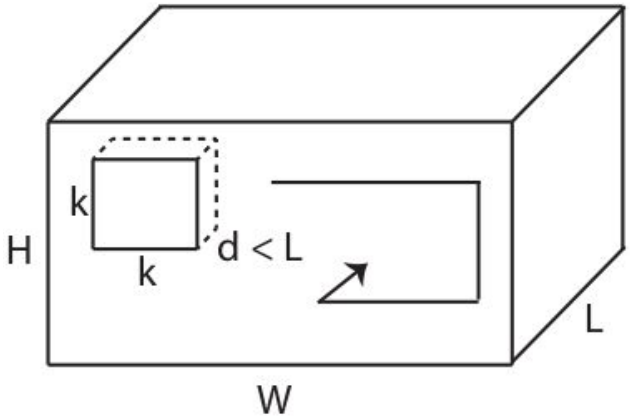
CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN
Sequence of clips	3D CNN	-	Pooling

3D CNN (C3D)

We can add an extra dimension to standard CNNs:

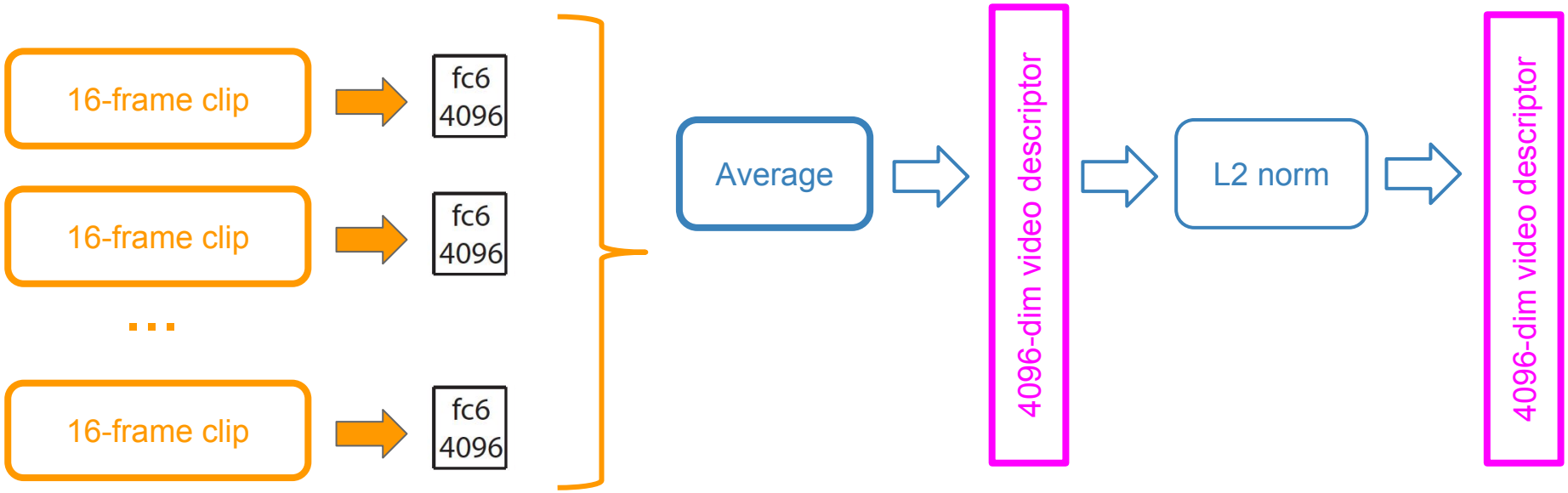
- An image is a HxWxD tensor: MxNxD’ conv filters
- A video is a TxHxWxD tensor: KxMxNxD’ conv filters



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

3D CNN (C3D)

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.

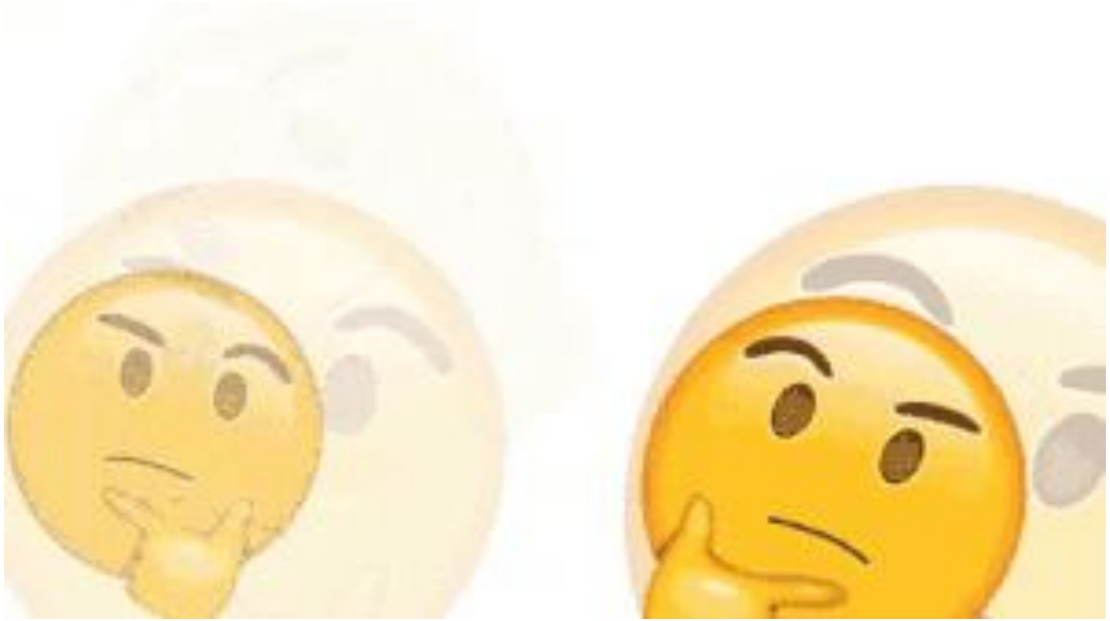


Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3D convolutional networks" ICCV 2015

3D CNN

Limitations:

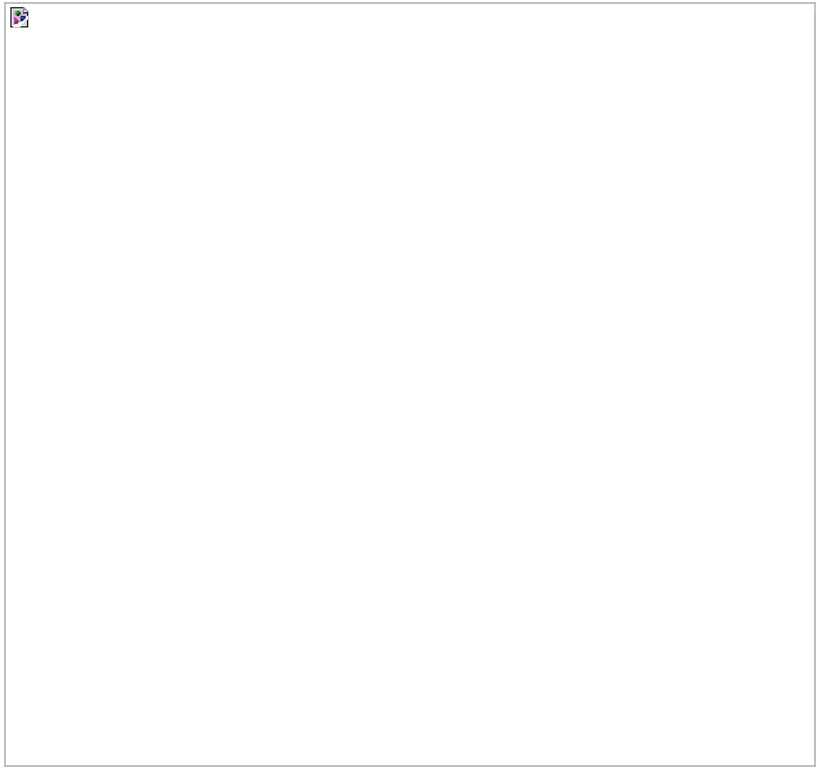
- How can we handle longer videos?
- How can we capture longer temporal dependencies?



CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN
Sequence of clips	3D CNN	-	Pooling
Sequence of clips	3D CNN	-	RNN

3D CNN + RNN



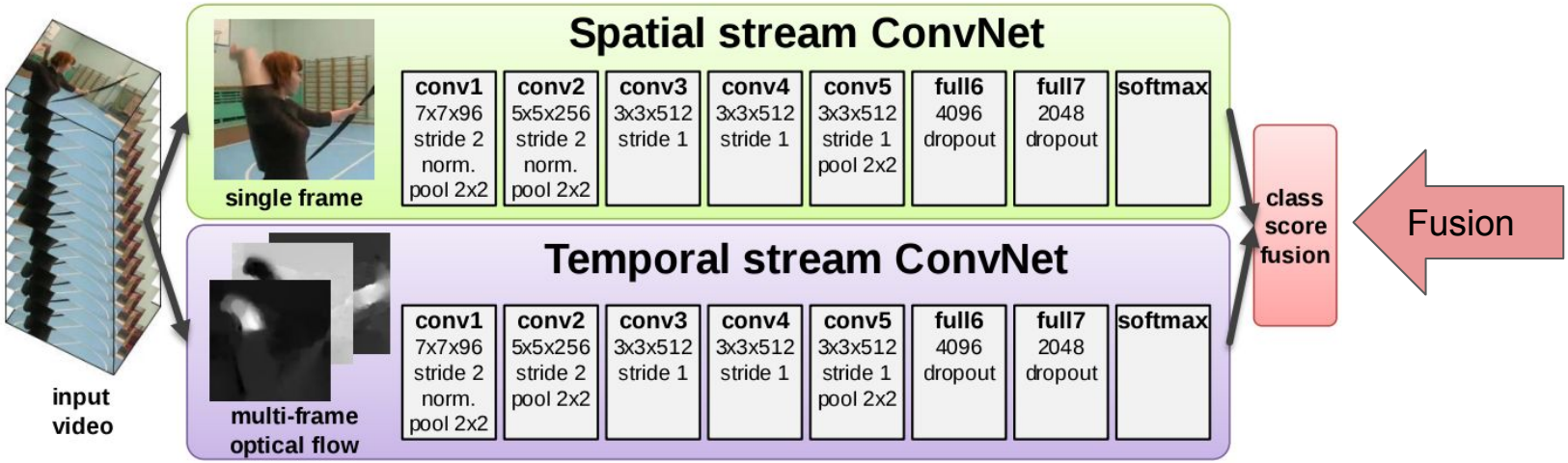
A. Montes, Salvador, A., Pascual-deLaPuente, S., and Giró-i-Nieto, X., "**Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks**", *NIPS Workshop 2016* **(best poster award)**

CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN
Sequence of clips	3D CNN	-	Pooling
Sequence of clips	3D CNN	-	RNN
Two-stream	2D CNN	2D CNN	Pooling

Two-streams 2D CNNs

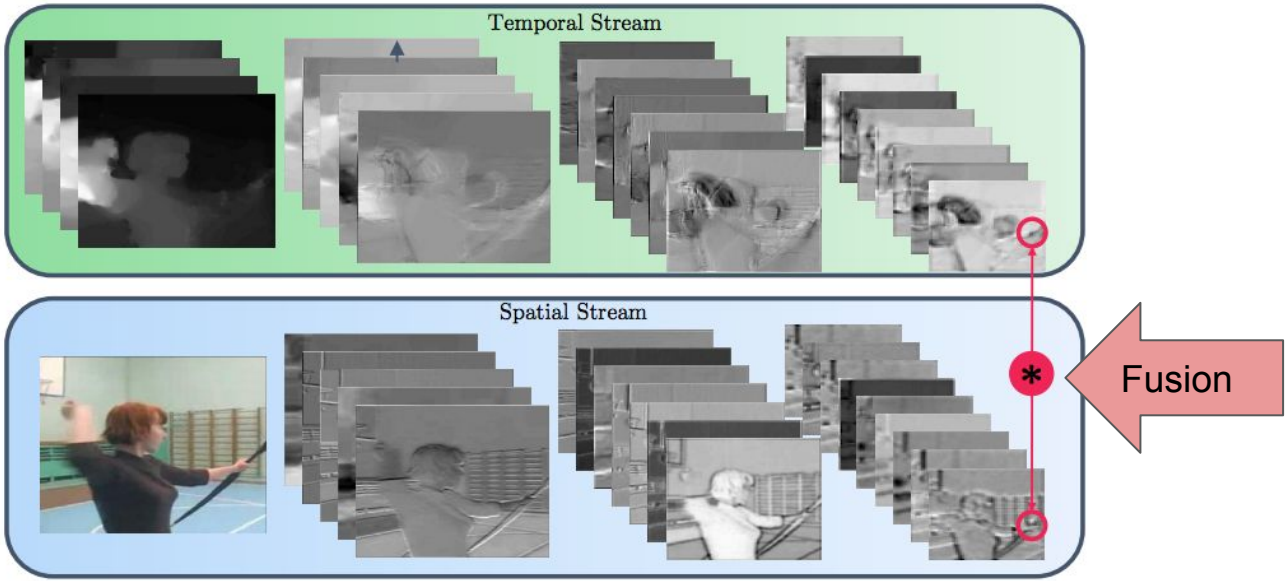
Problem: Single frame models do not take into account motion in videos.
Solution: extract optical flow for a stack of frames and use it as an input to a CNN.



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

Two-streams 2D CNNs

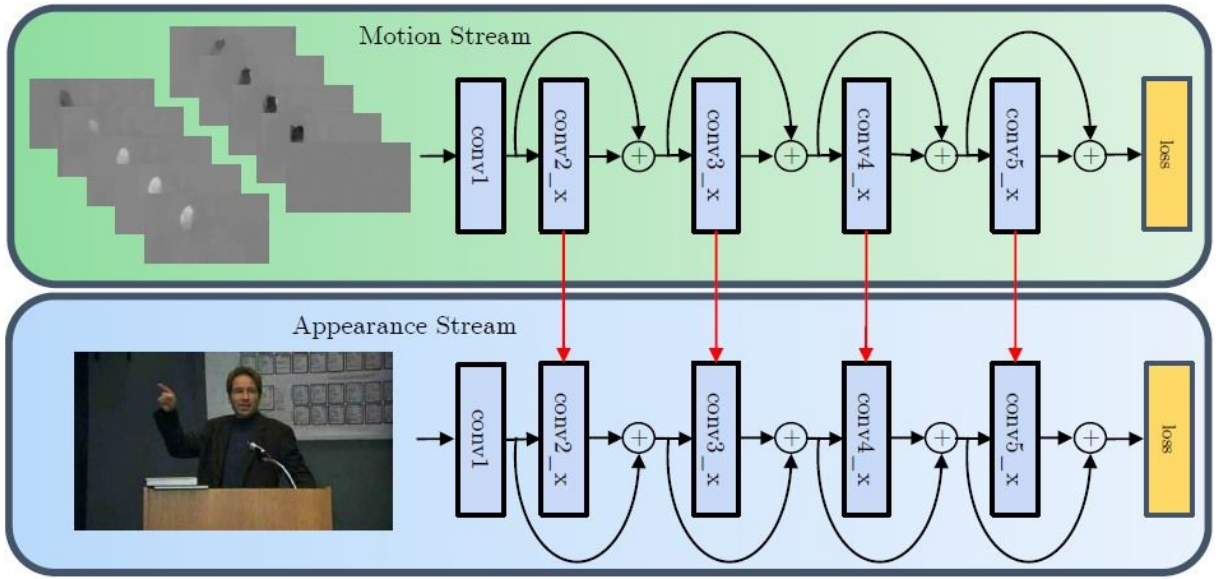
Problem: Single frame models do not take into account motion in videos.
Solution: extract optical flow for a stack of frames and use it as an input to a CNN.



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

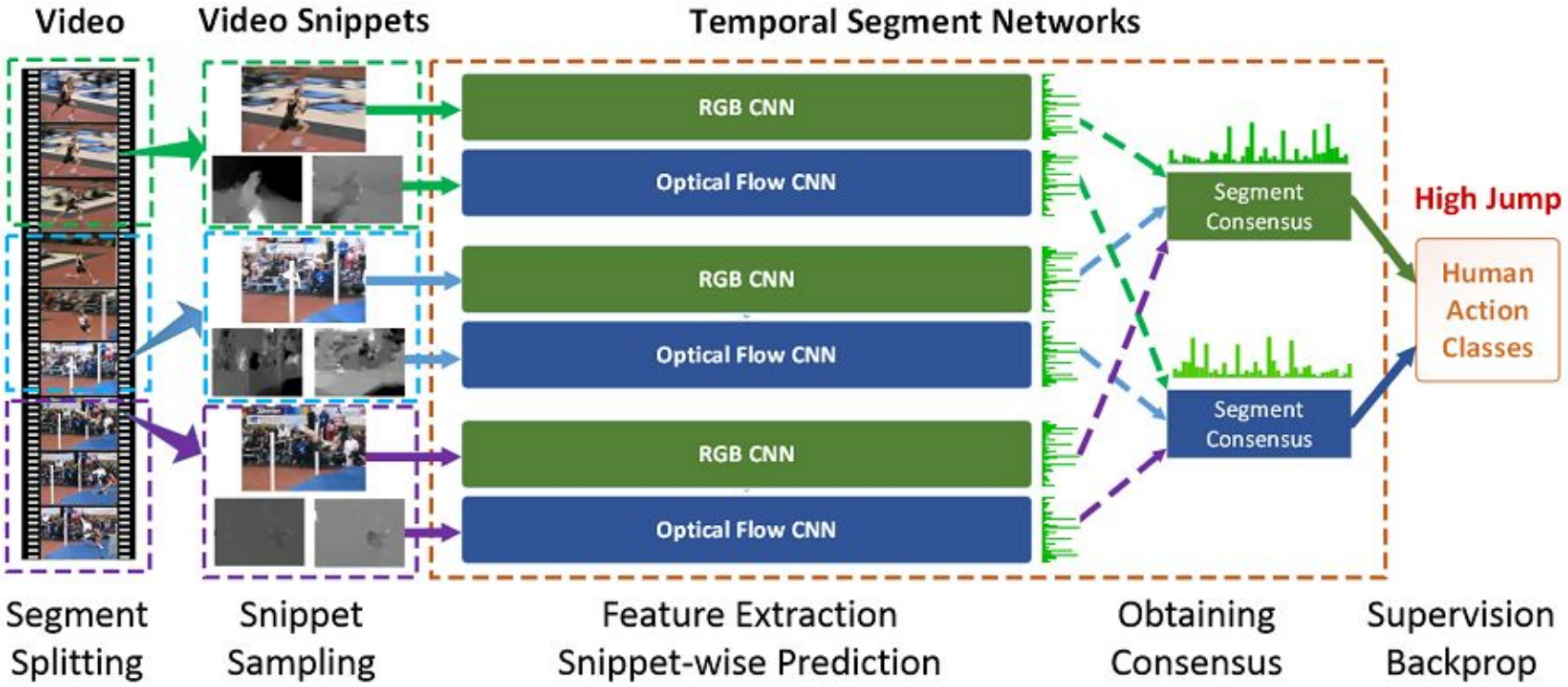
Two-streams 2D CNNs

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.



Feichtenhofer, Christoph, Axel Pinz, and Richard Wildes. "Spatiotemporal residual networks for video action recognition." NIPS 2016. [code]

Two-streams 2D CNNs



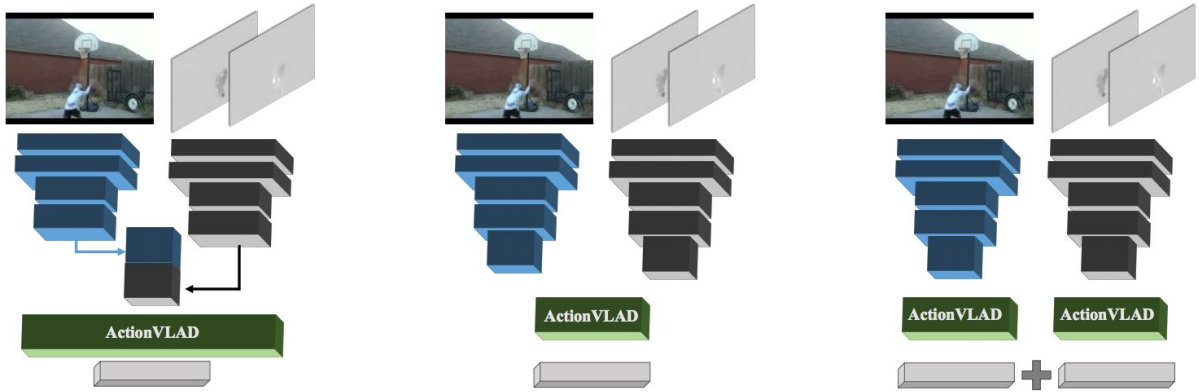
Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. "Temporal segment networks: Towards good practices for deep action recognition." ECCV 2016.

Two-streams 2D CNNs

Effect of training (rgb/flow)

Two-Stream	VLAD	ActionVLAD
47.1/55.2	44.9/55.6	51.2/58.4

How to fuse RGB and Flow?



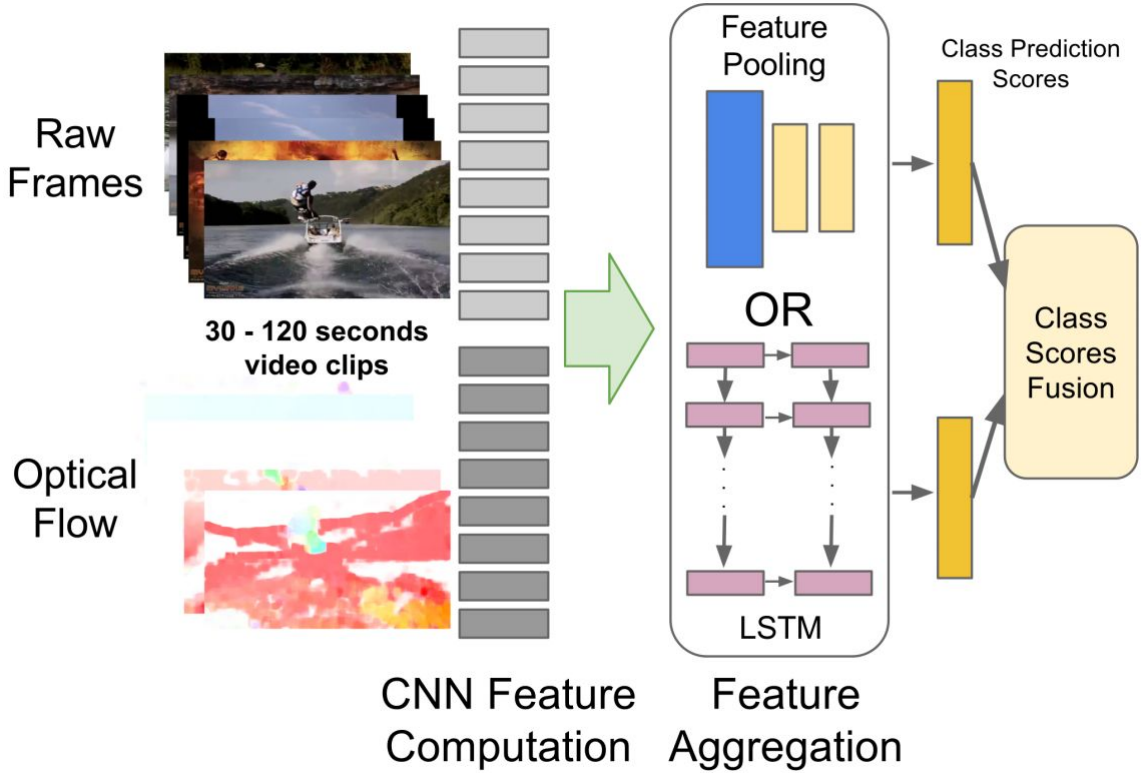
Concat Fuse	Early Fuse	Late Fuse
56.0	64.8	66.9

Girdhar, Rohit, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. "ActionVLAD: Learning spatio-temporal aggregation for action classification." CVPR 2017.

CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN
Sequence of clips	3D CNN	-	Pooling
Sequence of clips	3D CNN	-	RNN
Two-stream	2D CNN	2D CNN	Pooling
Two-stream	2D CNN	2D CNN	RNN

Two-streams 2D CNNs + RNN

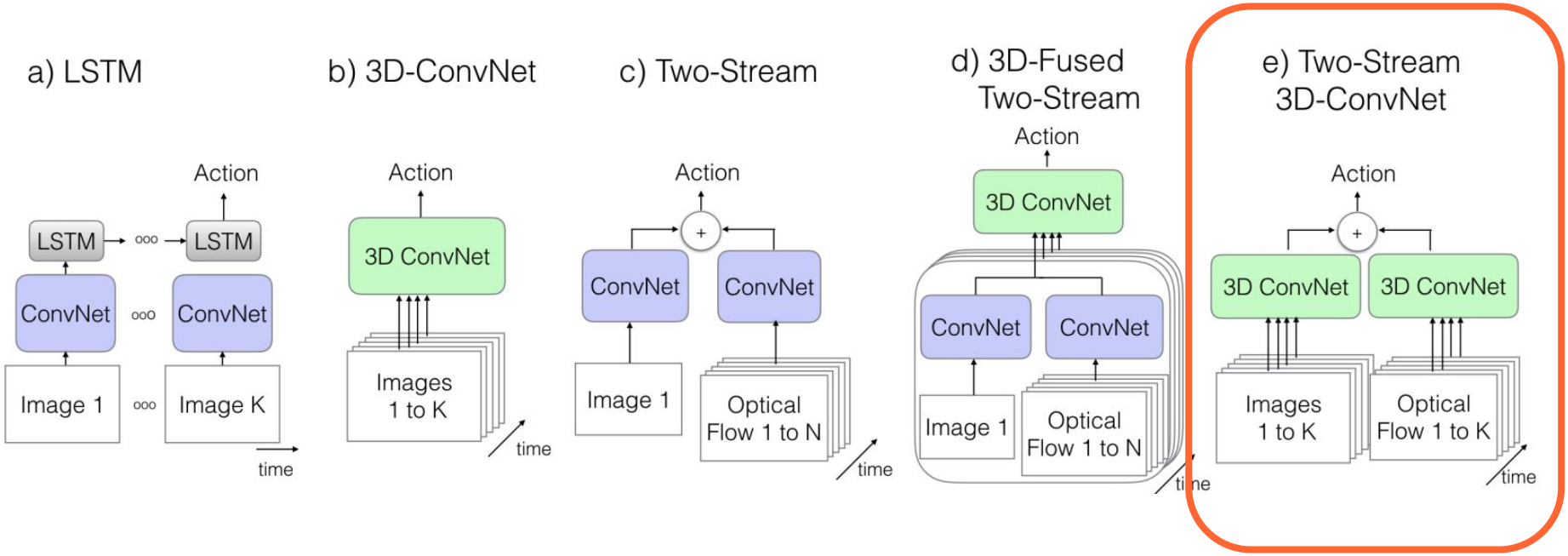


Yue-Hei Ng, Joe, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. "Beyond short snippets: Deep networks for video classification." CVPR 2015

CNNs for sequences of images

CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN
Sequence of clips	3D CNN	-	Pooling
Sequence of clips	3D CNN	-	RNN
Two-stream	2D CNN	2D CNN	Pooling
Two-stream	2D CNN	2D CNN	RNN
Two-stream	Inflated 3D CNN	Inflated 3D CNN	Pooling

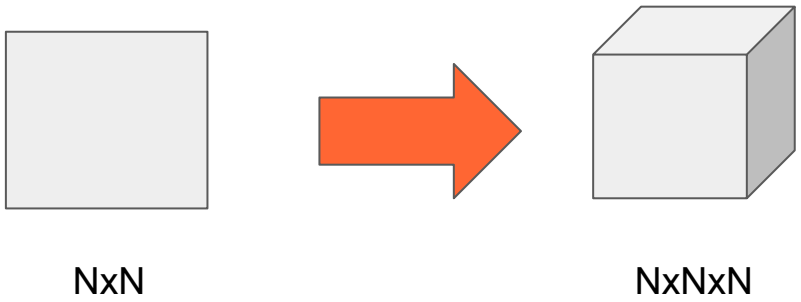
Two-streams 3D CNNs



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

Two-streams Inflated 3D CNNs (I3D)

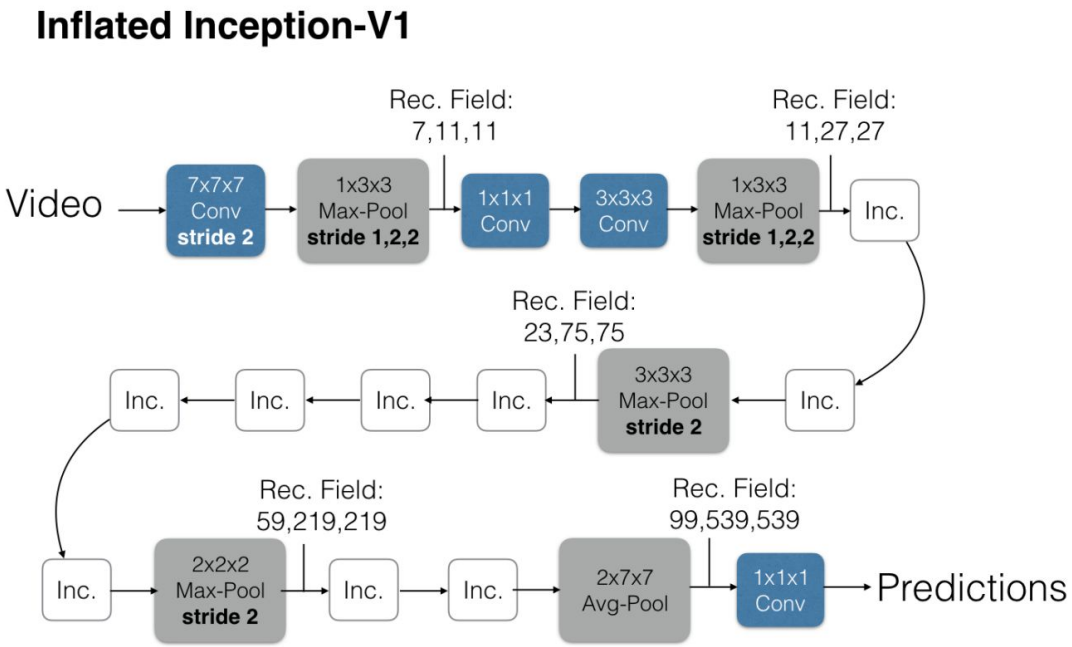
Adapt 2D CNNs found for ImageNet classification to 3D convolutions

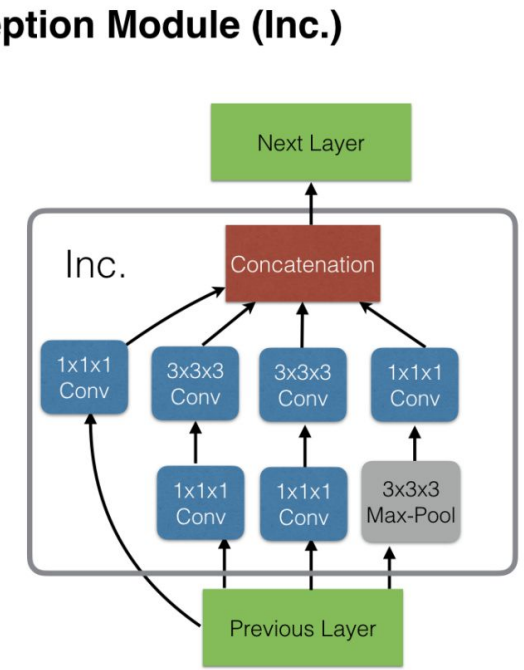


3D models are initialized with ImageNet images transformed into ‘boring’ video sequences.

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

Two-streams 3D CNNs





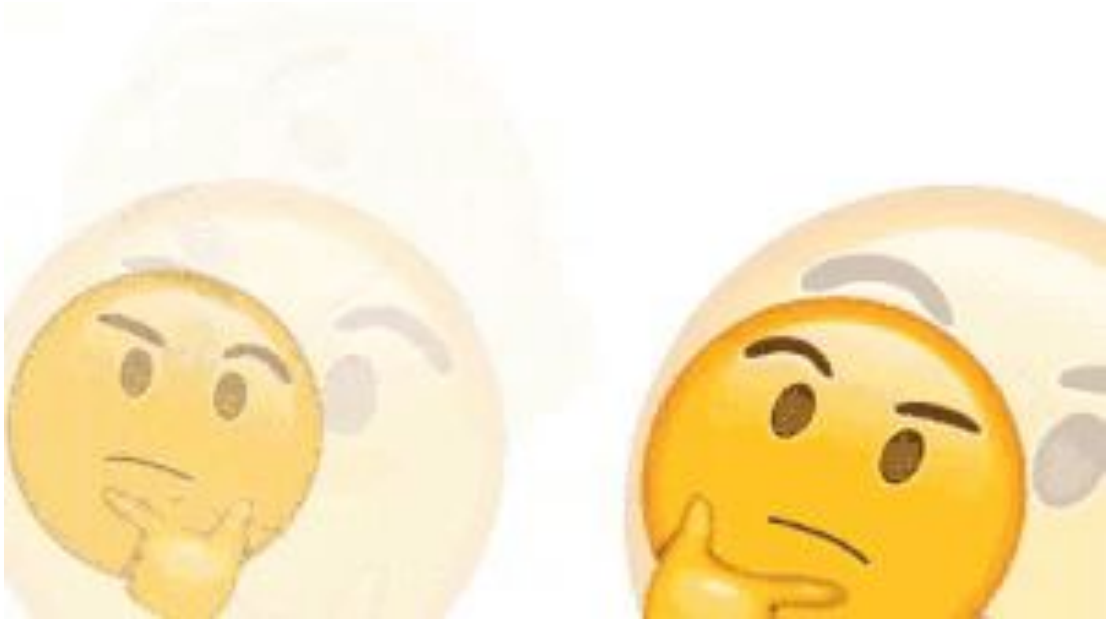
Carreira, J., & Zisserman, A. . Quo vadis, action recognition? A new model and the kinetics dataset. CVPR 2017. [code]

CNNs for sequences of images

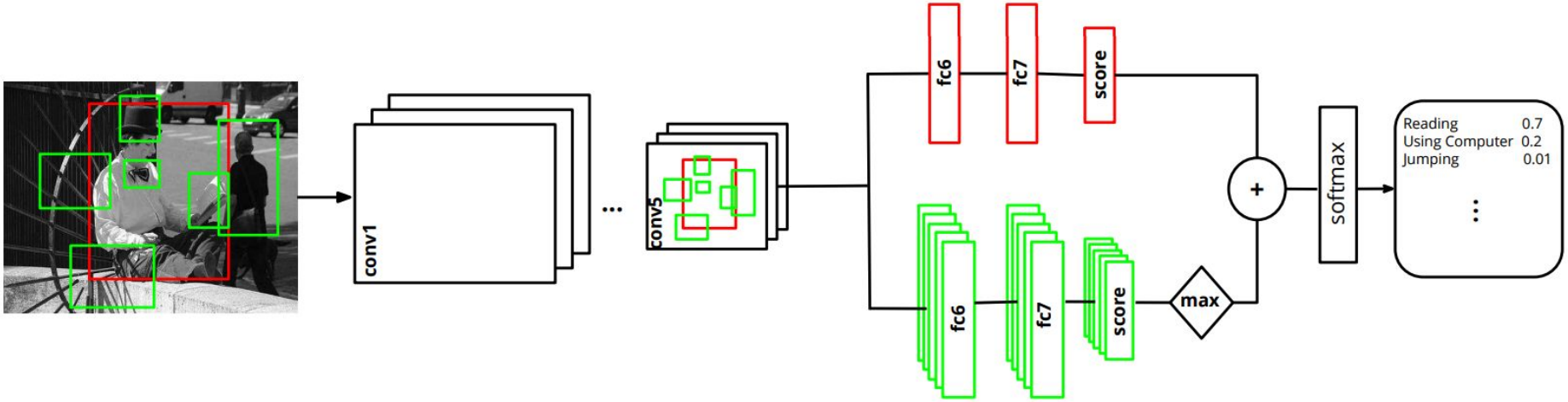
CNN Input	RGB	Optical Flow	Fusion
Single frame	2D CNN	-	Pooling + NN
Multiple frames	2D CNN	-	Pooling + NN
Sequence of images	2D CNN	-	RNN
Sequence of clips	3D CNN	-	Pooling
Sequence of clips	3D CNN	-	RNN
Two-stream	2D CNN	2D CNN	Pooling
Two-stream	2D CNN	2D CNN	RNN
Two-stream	Inflated 3D CNN	Inflated 3D CNN	Pooling

Action recognition

Which deep learning techniques at a local scale may help in action recognition ?

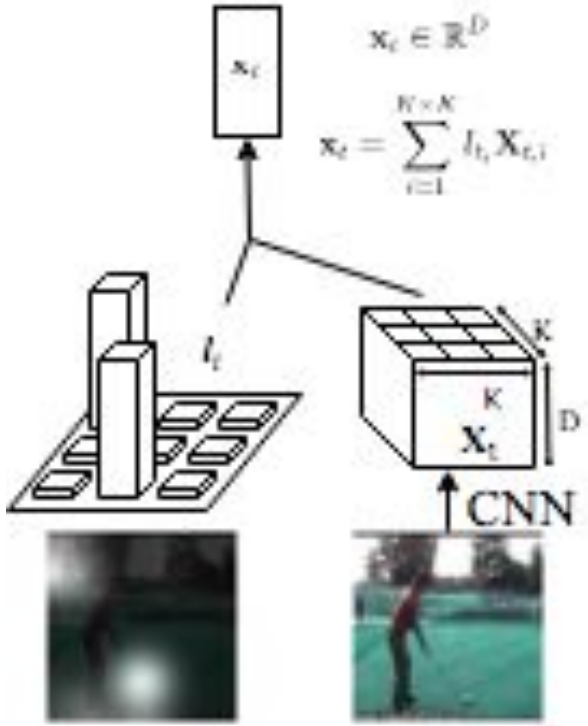


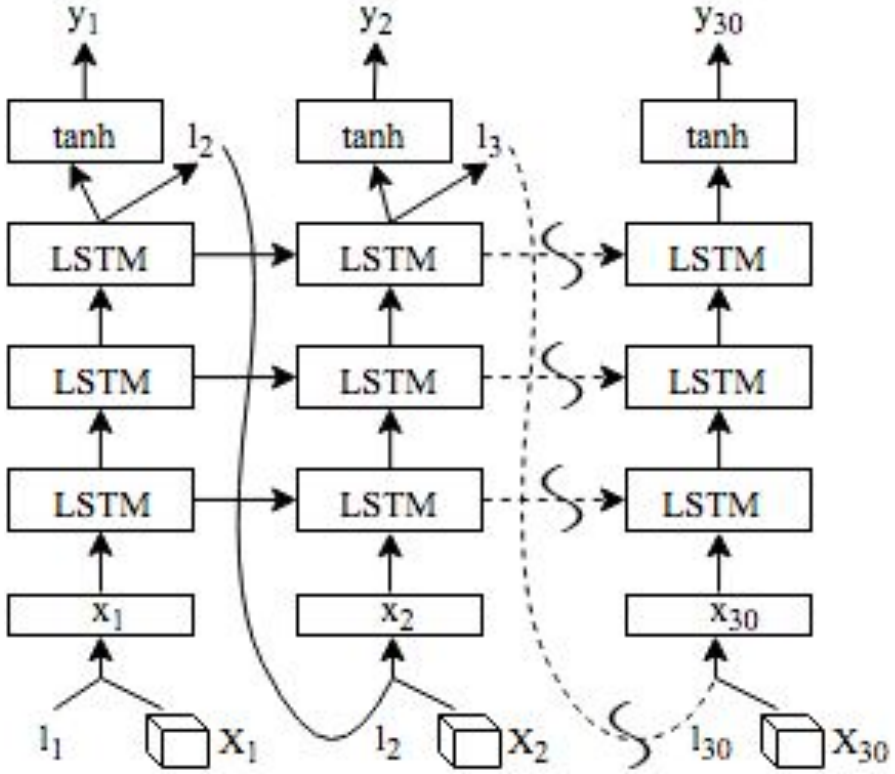
Action Recognition with object detection



Gkioxari, Georgia, Ross Girshick, and Jitendra Malik. "Contextual action recognition with r* cnn." In ICCV 2015. [code]

Action Recognition with attention



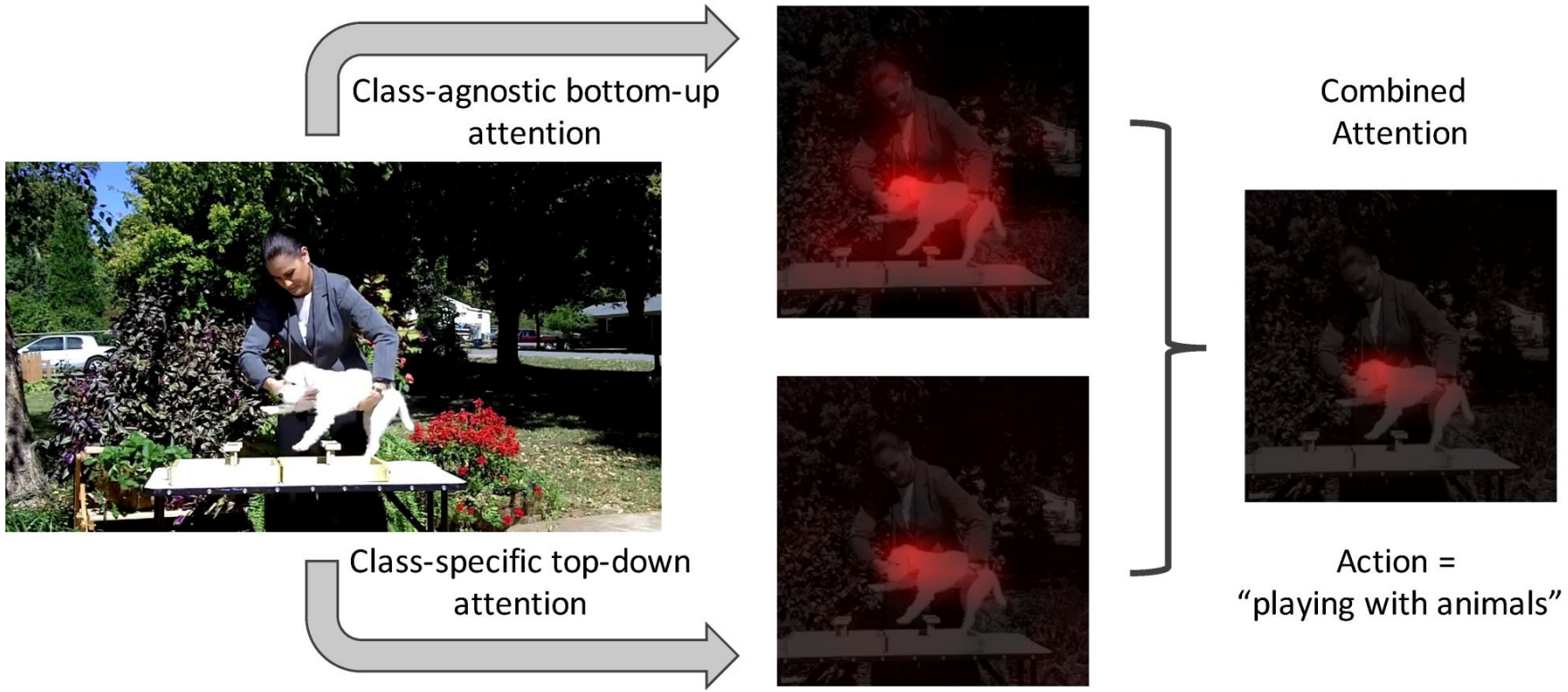


Sharma, Shikhar, Ryan Kiros, and Ruslan Salakhutdinov. "Action recognition using visual attention." ICLRW 2016.

Action Recognition with soft attention

Sharma, Shikhar, Ryan Kiros, and Ruslan Salakhutdinov. "Action recognition using visual attention." ICLRW 2016.

Action recognition with soft attention



Girdhar, Rohit, and Deva Ramanan. "Attentional pooling for action recognition." NIPS 2017

Action Recognition with hard attention

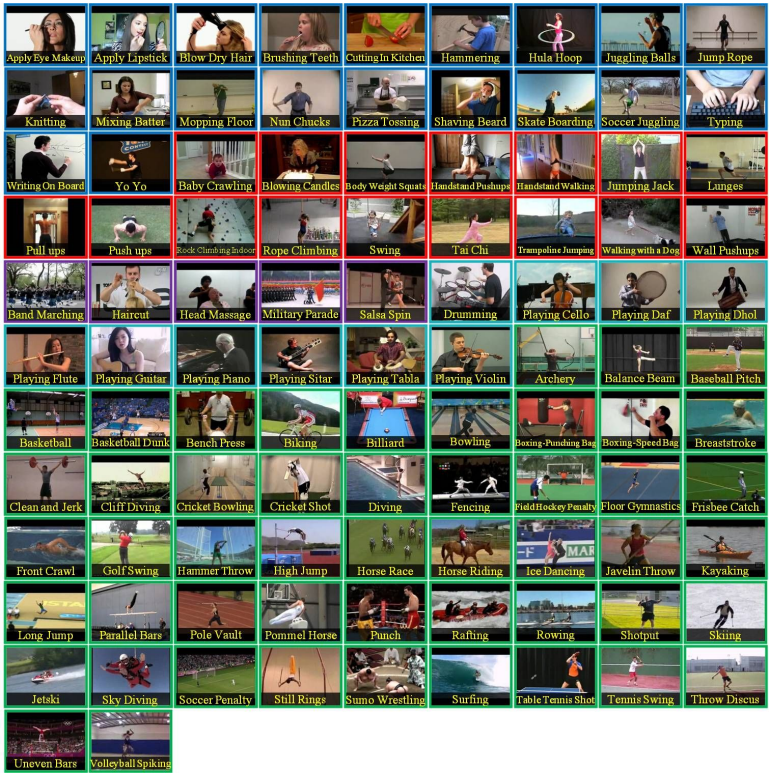
Figure 1. Key volumes detected by our key volume mining deep framework. A volume is a spatial-temporal video clip. The top row shows key volumes are very sparse among the whole video, and the second row shows that key volumes may come from different modalities (different motion patterns here). Note that frames are sampled with fixed time interval.

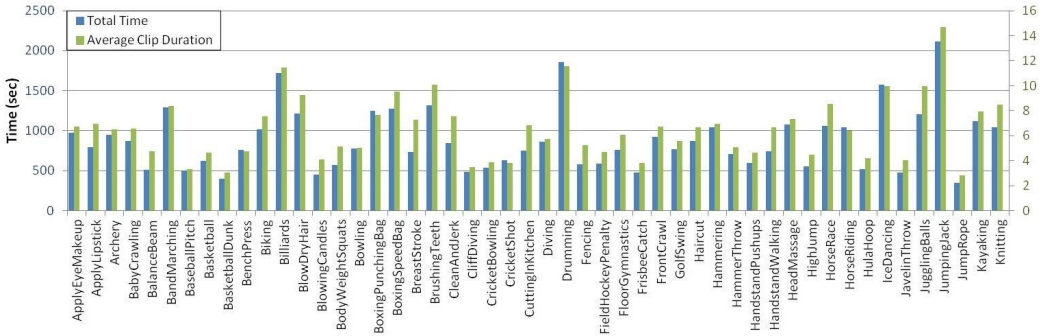
Zhu, Wangjiang, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. "A key volume mining deep framework for action recognition." CVPR 2016.

Outline

1. Architectures
2. **Datasets**
3. Tips and tricks

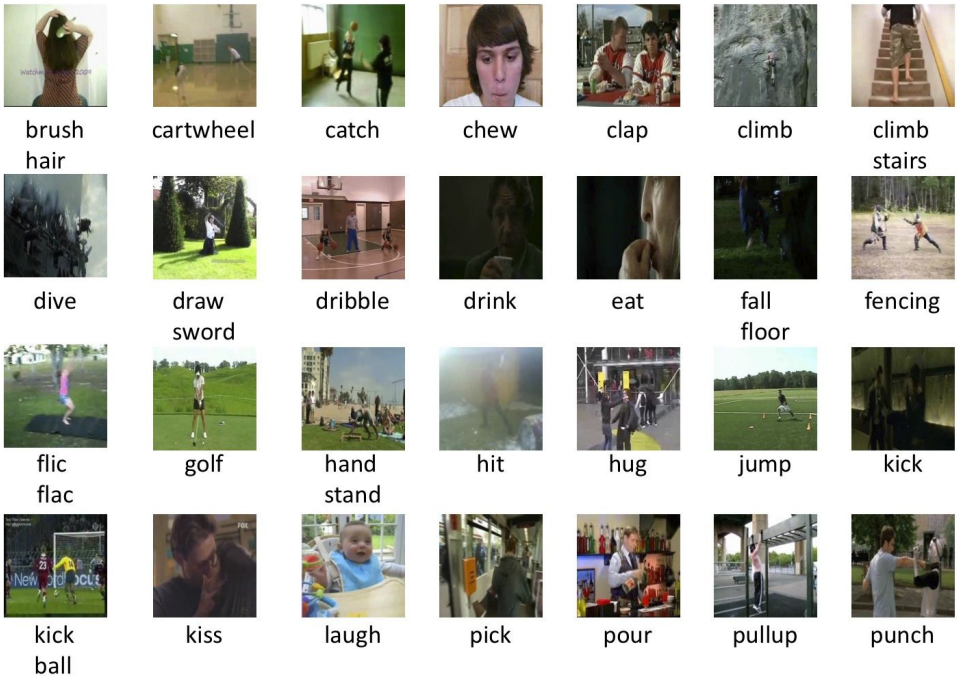
Datasets: UCF-101





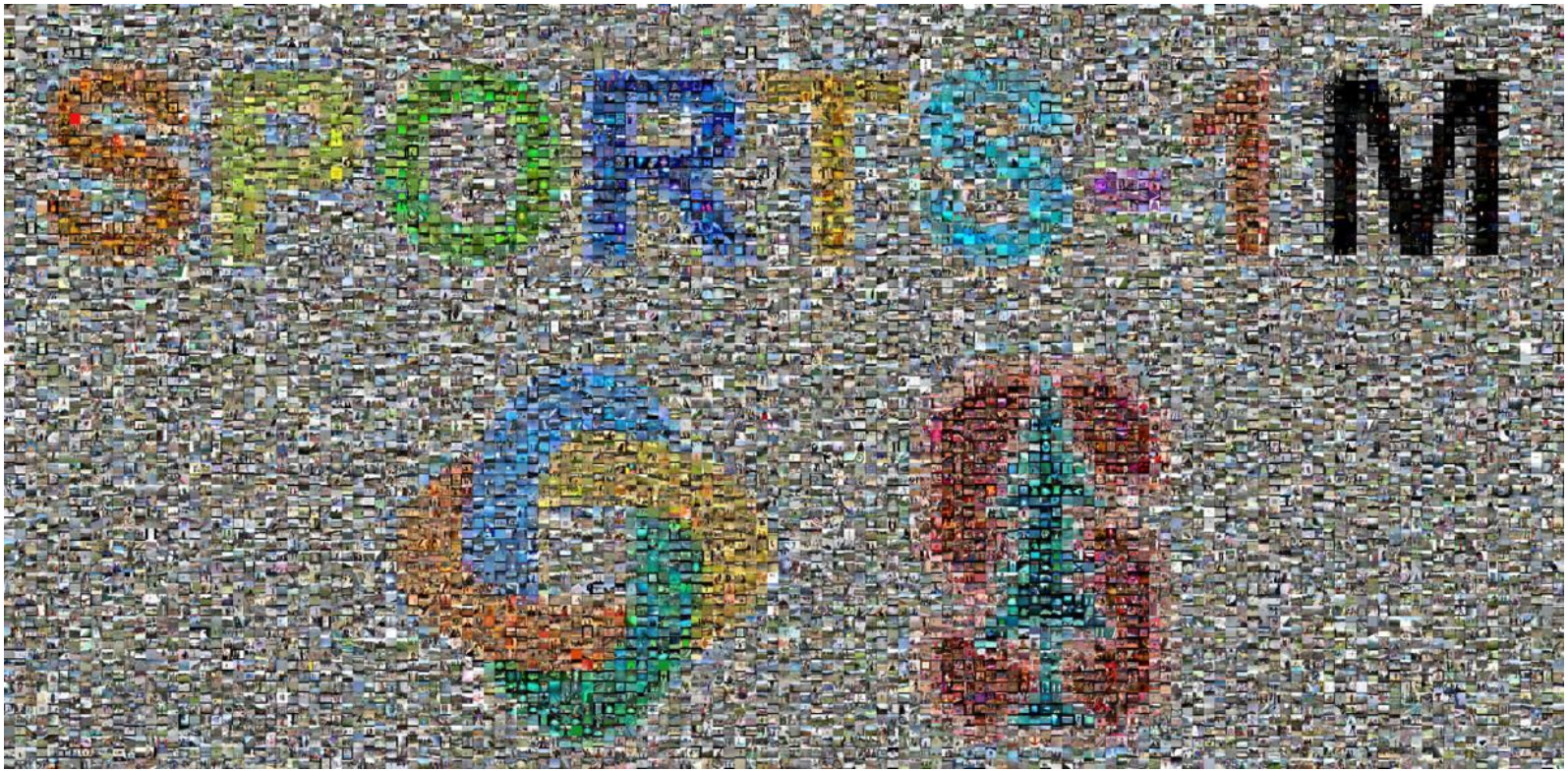
Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

Datasets: HDMB51 (Brown University)



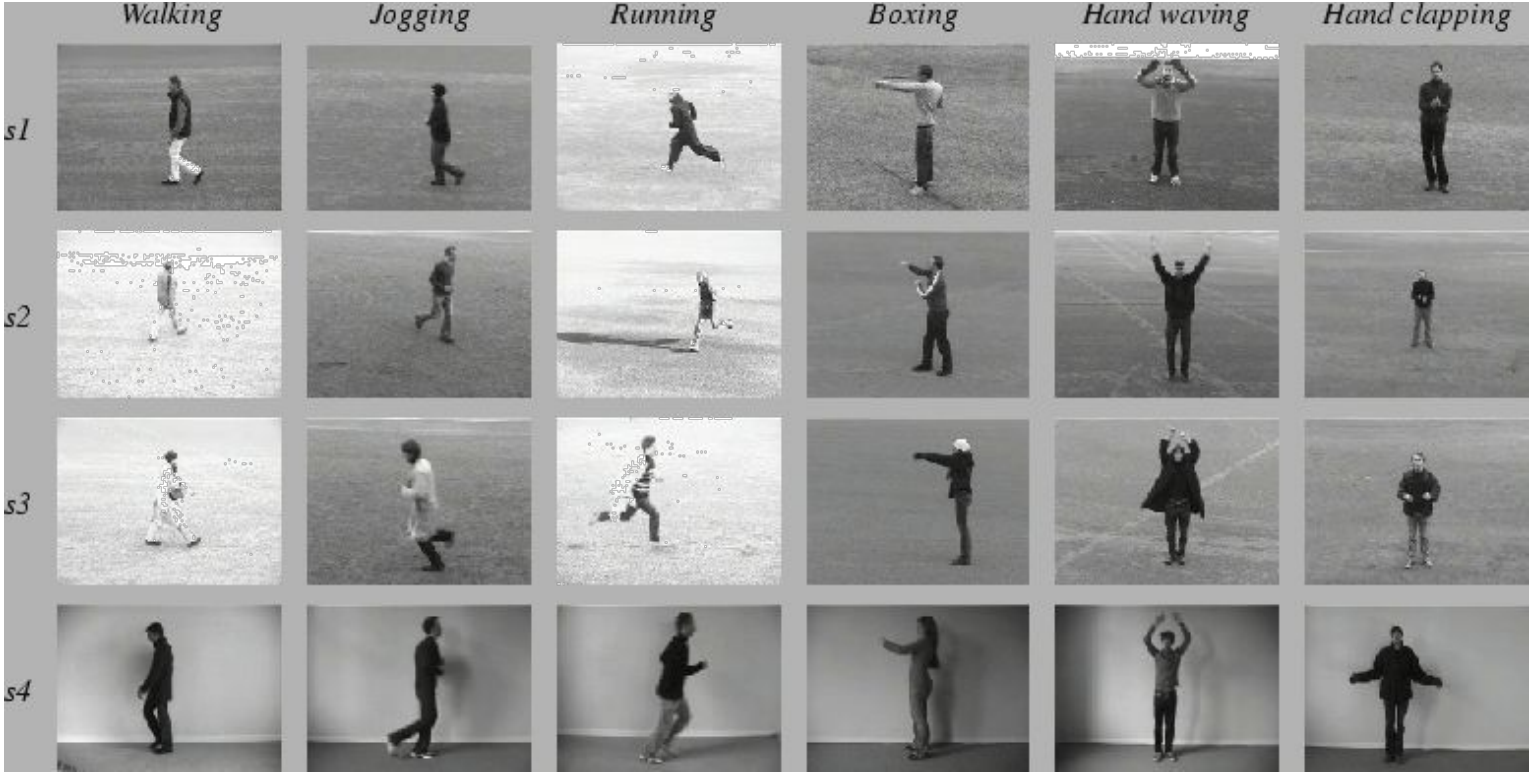
Kuehne, Hildegard, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. "HMDB: a large video database for human motion recognition." ICCV 2011.

Datasets: Sports 1M (Stanford)



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

Datasets: KTH

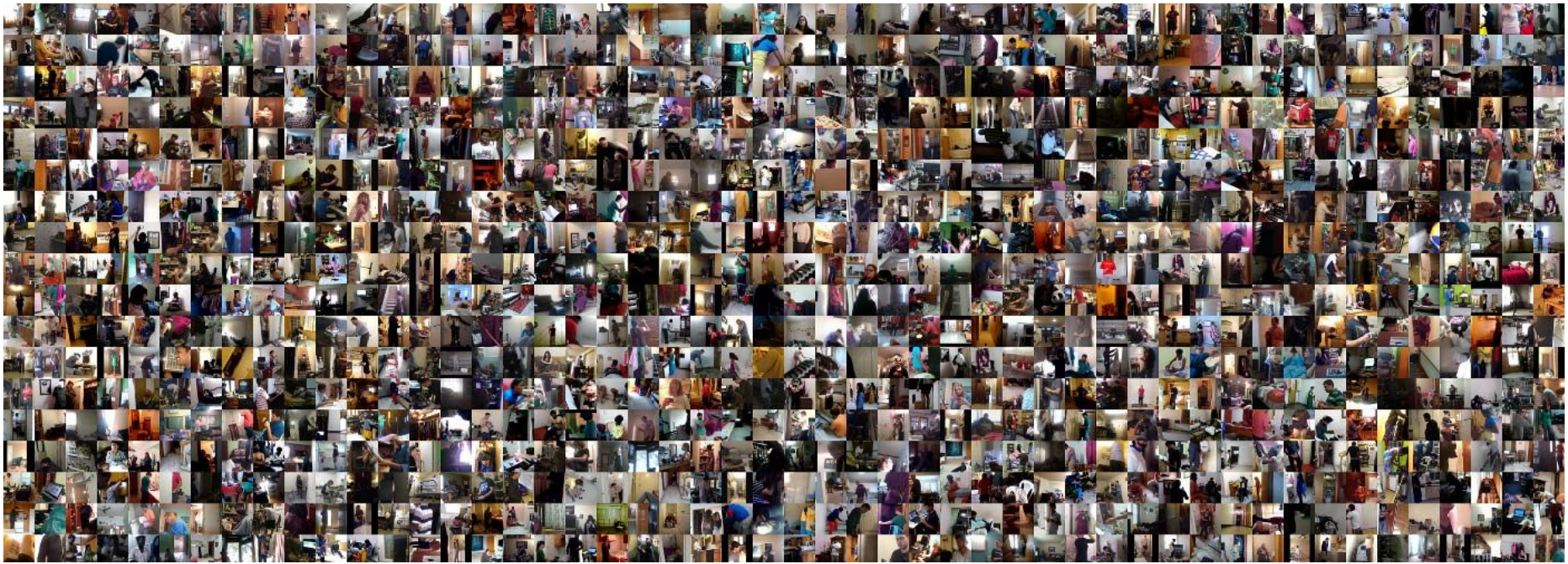


Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." In Pattern Recognition, 2004. ICPR 2004.

Datasets: ActivityNet

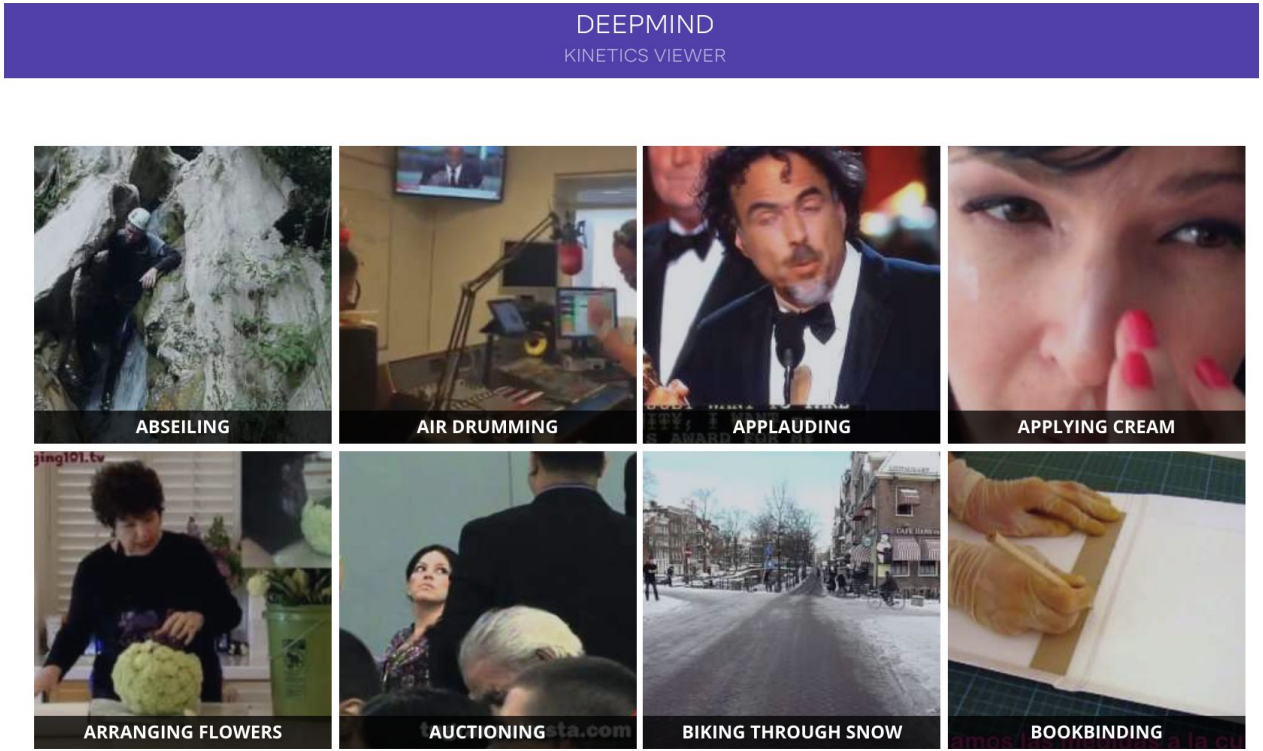
Heilbron, F.C., Escorcia, V., Ghanem, B. and Niebles, J.C.,. "Activitynet: A large-scale video benchmark for human activity understanding". CVPR 2015.

Datasets: Charades (Allen AI)



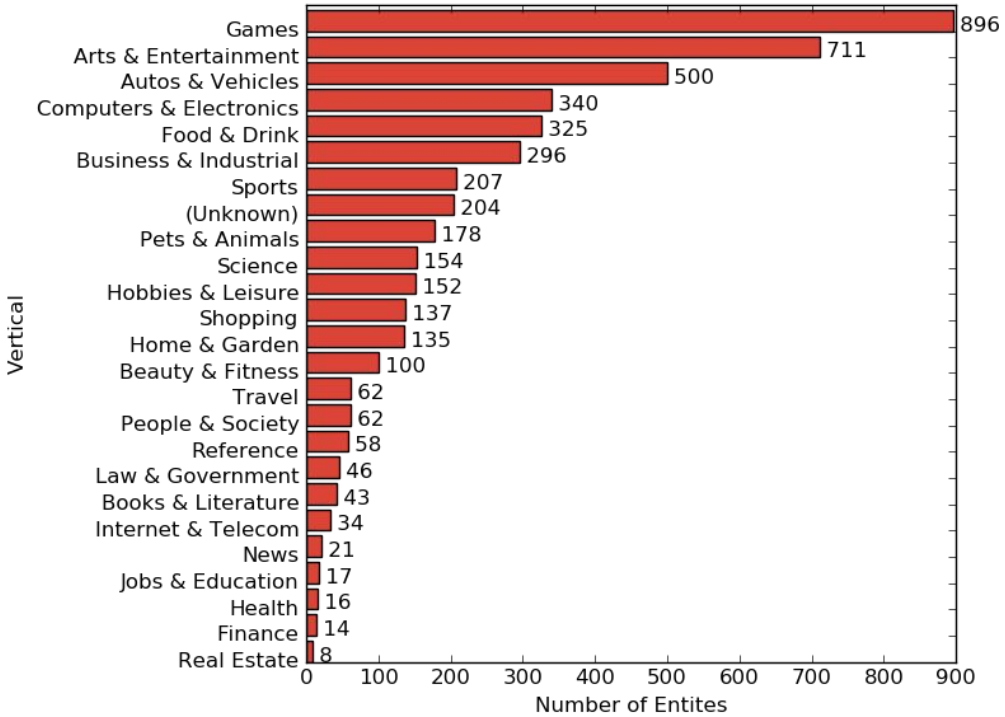
Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016, October). Hollywood in homes: Crowdsourcing data collection for activity understanding. ECCV 2016. [Dataset] [Code]

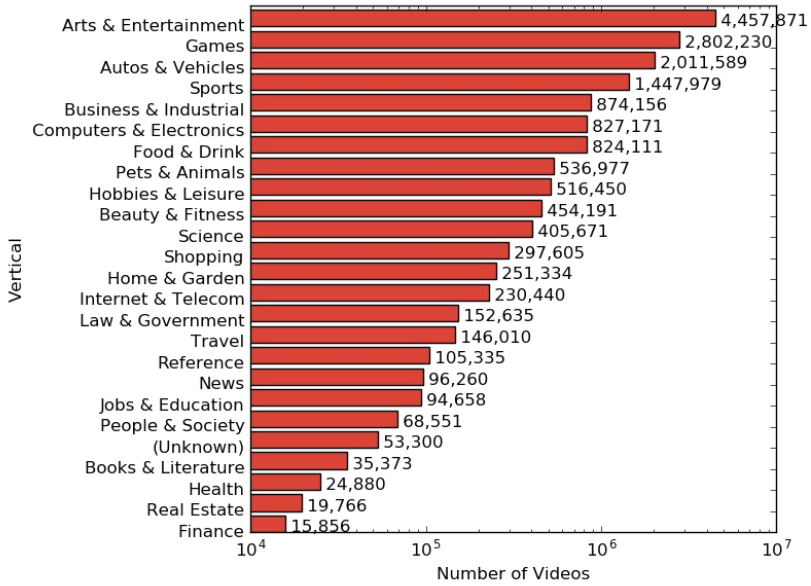
Datasets: Kinectics (DeepMind)



Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Suleyman, M. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

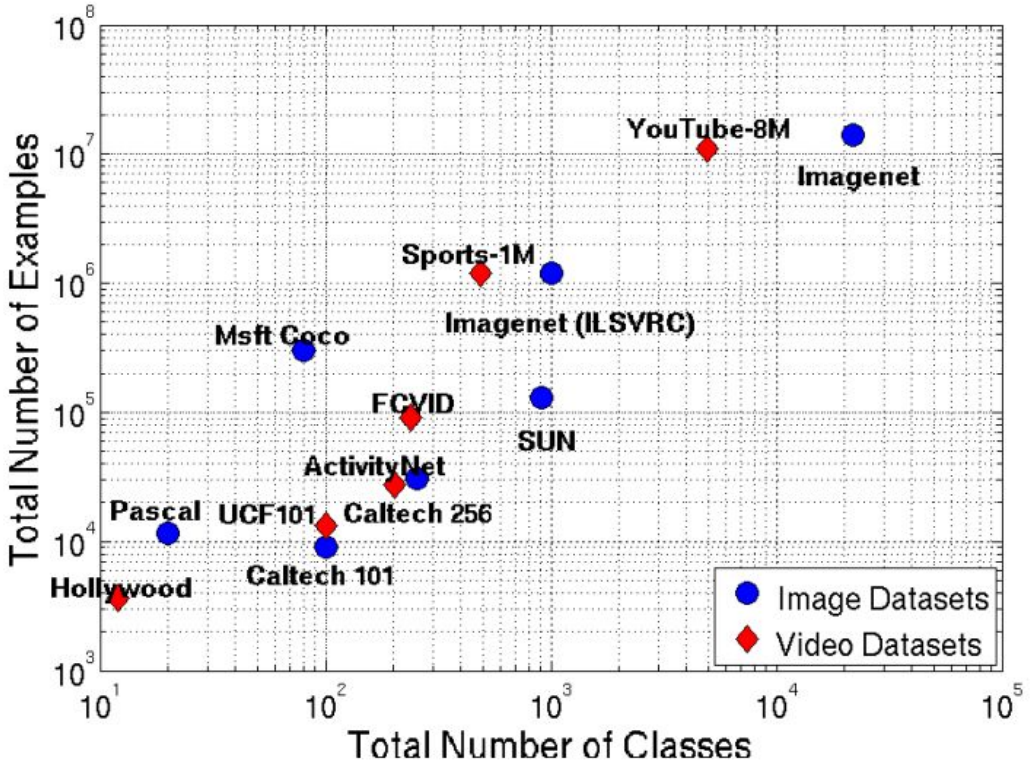
Datasets: YouTube-8M (Google)





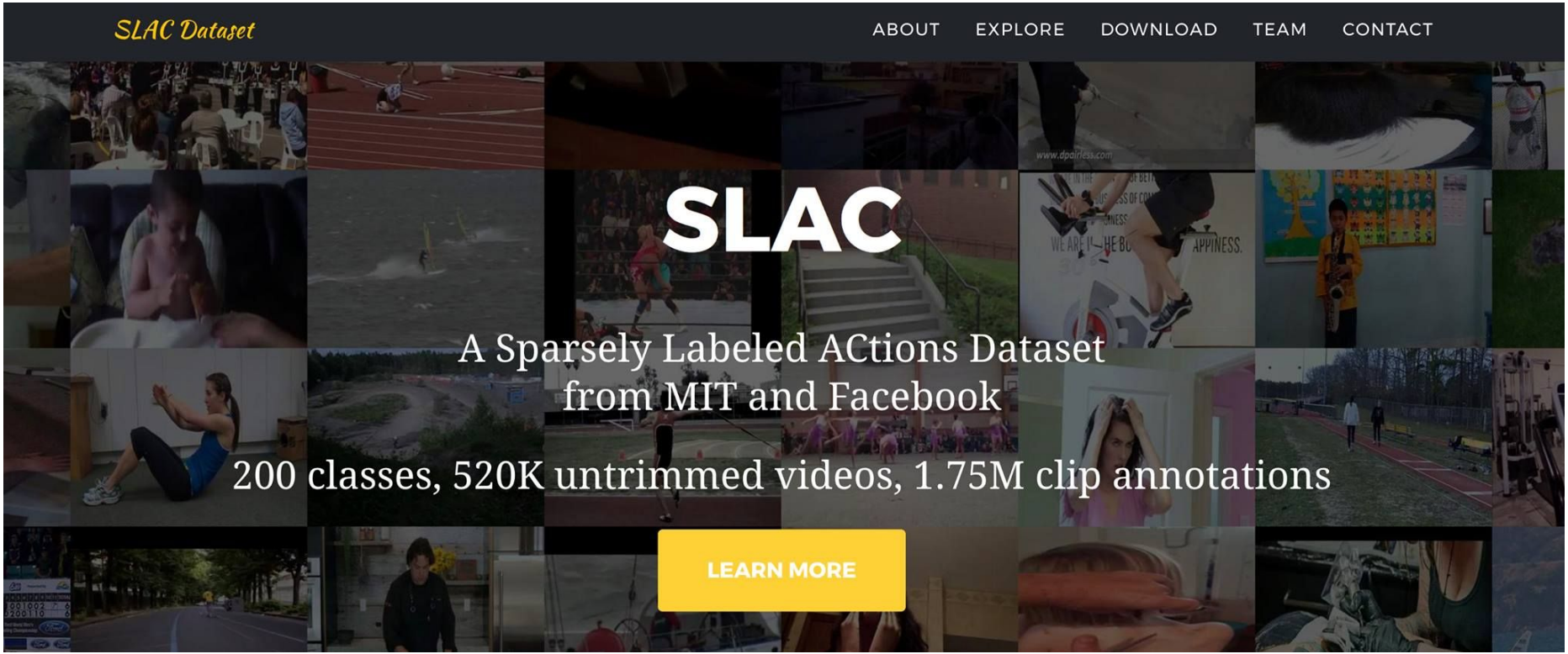
(Slides by Dídac Surís) Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. "Youtube-8m: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675 (2016). [project]

Activity Recognition: Datasets



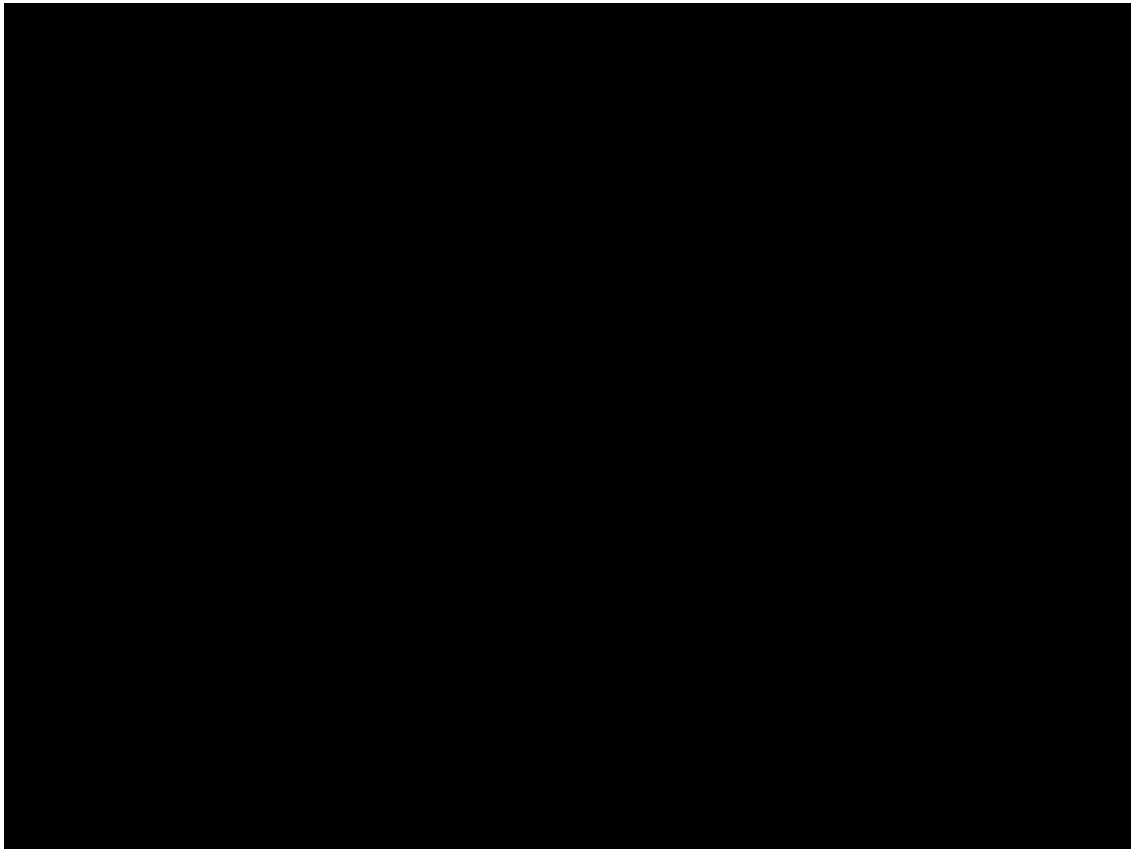
(Slides by Dídac Surís) Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. "Youtube-8m: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675 (2016). [project]

Datasets: SLAC (MIT & Facebook)



Hang Zhao, Zhicheng Yan, Heng Wang, Lorenzo Torresani, Antonio Torralba, "SLAC: A Sparsely Labeled Dataset for Action Classification and Localization" arXiv 2017 [project page]

Datasets: Moments in Time (MIT & IBM)



Monfort, Mathew, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown et al. "Moments in Time Dataset: one million videos for event understanding." arXiv preprint arXiv:1801.03150 (2018).

Datasets: DALY (INRIA)

DALY contains the following spatial annotations:

- bounding box around the action
- upper body pose annotation, including a bounding box around the head
- bounding box around object(s) involved in the action

Weinzaepfel, Philippe, Xavier Martin, and Cordelia Schmid. "Human Action Localization with Sparse Spatial Supervision." (2017).

Datasets: AVA (Berkeley & Google)

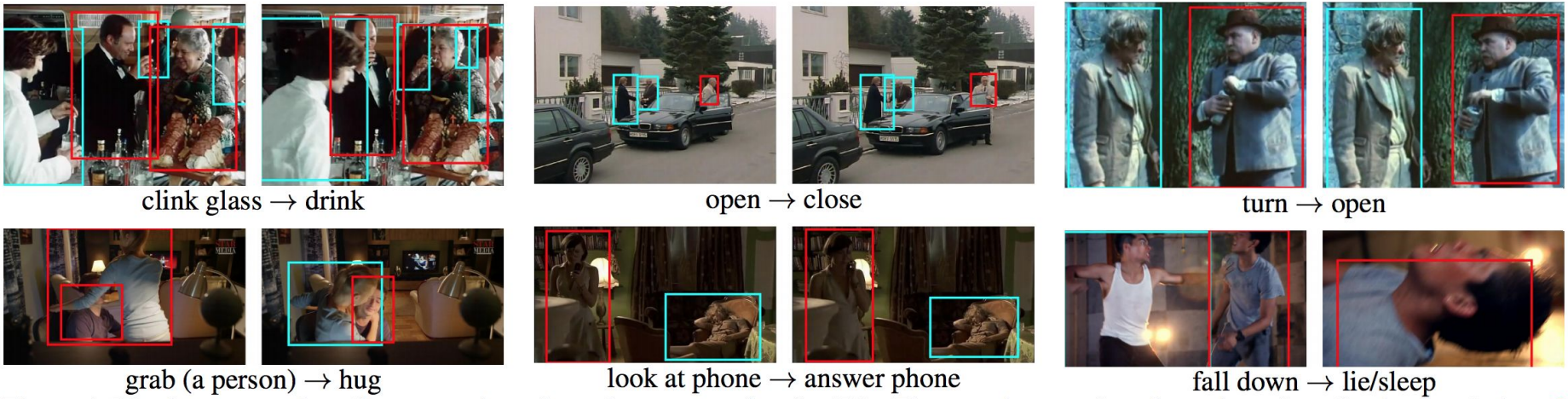


Figure 4. We show examples of how atomic actions change over time in AVA. The text shows pairs of atomic actions for the people in red bounding boxes. Temporal information is key for recognizing many of the actions and appearance can substantially vary within an action category, such as opening a door or bottle.

Gu, C., Sun, C., Vijayanarasimhan, S., Pantofaru, C., Ross, D. A., Toderici, G., ... & Malik, J. (2017). AVA: A video dataset of spatio-temporally localized atomic visual actions. arXiv preprint arXiv:1705.08421.

Outline

1. Architectures
2. Datasets
3. **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

Tips & Tricks by Víctor Campos (2017)

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.

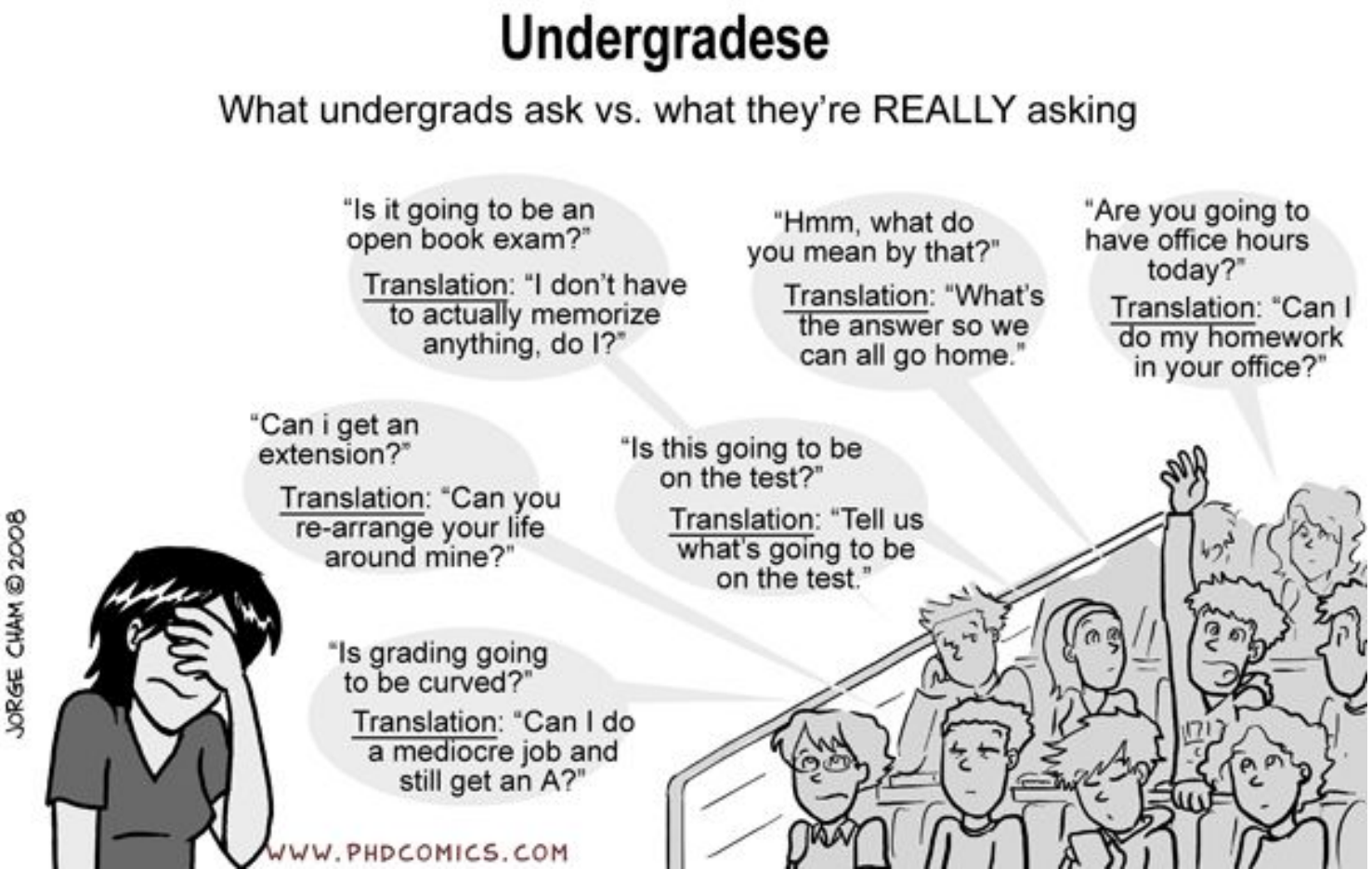
Tips & Tricks by Víctor Campos (2017)

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

Tips & Tricks by Víctor Campos (2017)

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
WWW.PHDCOMICS.COM

Deep Learning online courses by UPC:

- [MSc course](#) (2017)
- [BSc course](#) (2018)

Next edition Autumn 2018

- [1st edition](#) (2016)
- [2nd edition](#) (2017)
- [3rd edition](#) (2018)

Summer School (late June 2018)

- [1st edition](#) (2017)
- [2nd edition](#) (2018)

Next edition Winter/Spring 2019