Video analysis

Hakan Bilen

Machine Learning Practical - MLP Lecture 18
13 March 2019

IBM prize

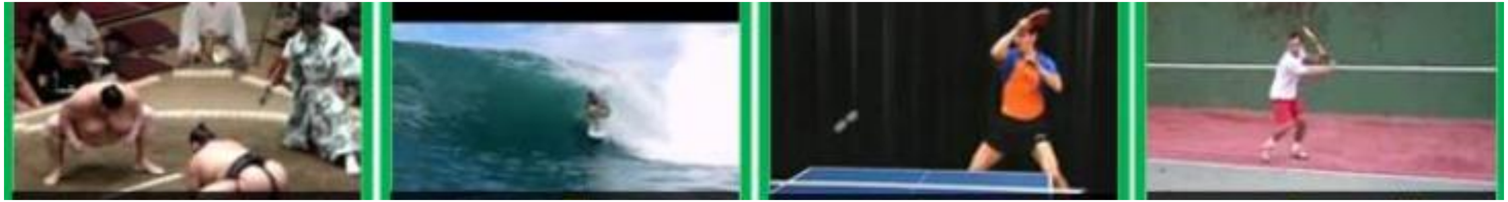
The Event:

- Friday 26 April, 12:30 followed by a reception
- 5-minute presentations from the short-listed projects
- Prizes awarded

The Process:

- Short-list constructed by MLP instructors based on final reports
- Short-list judged by a panel which will also include ML people not involved with the course

What actions are performed in the images?



a) Tennis swing
b) Table tennis shot
c) Sumo wrestling
d) Surfing

How about these ones?

a) Closing a laptop

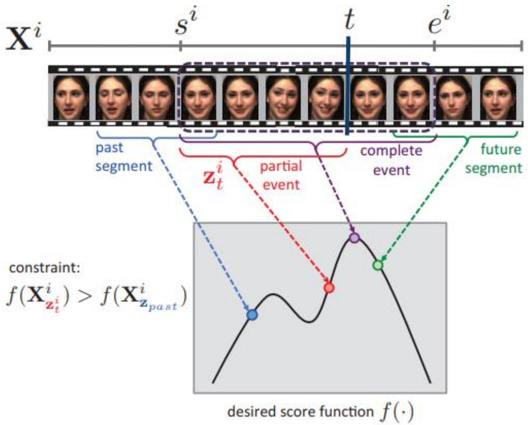
b) Opening a laptop

c) Putting down a laptop

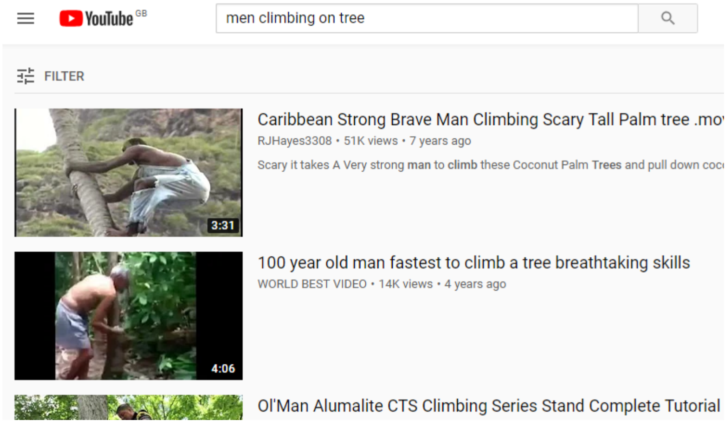
d) Taking a laptop

So far ...

- Images and sentences (sequences of words)
- CNNs, RNNs, GANs
- Today: Sequences of frames, **videos**!
- Video analysis
  - Action recognition (classification, detection)
  - Early event prediction
  - Video retrieval and captioning
  - Video summarization
  - ...



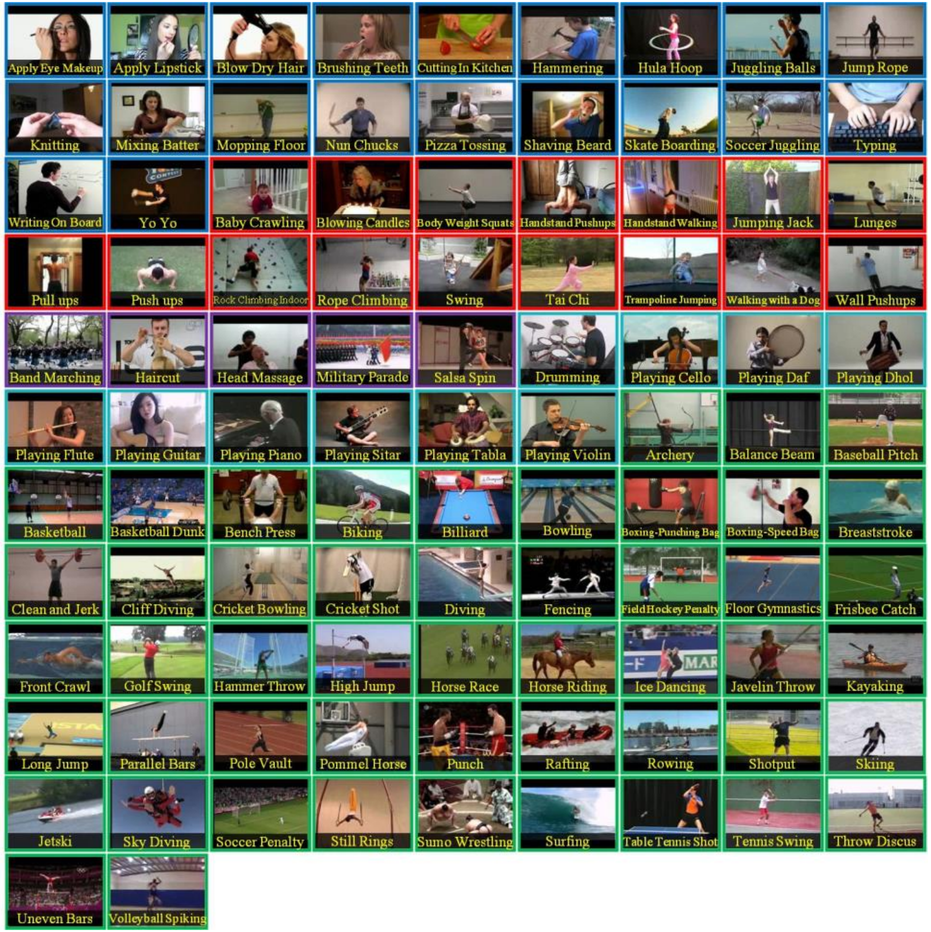
Credit: Hoai and De la Torre



Datasets – UCF101

- Realistic action videos, collected from YouTube
- 101 action categories
- 13320 videos
- 5 super categories:

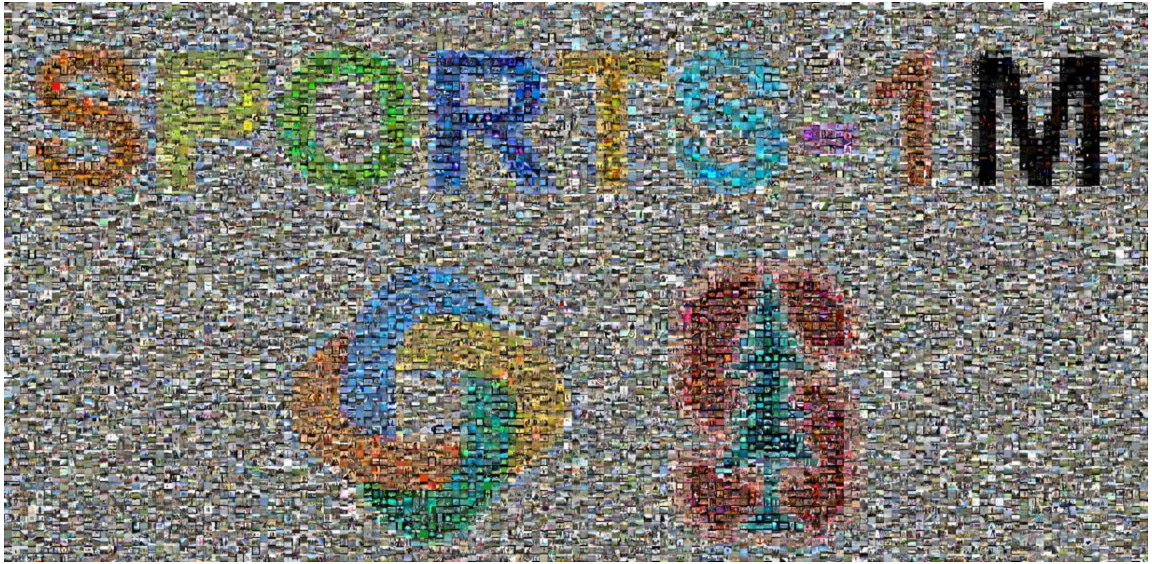
    - Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, Sports



[Soomro et al (2012), UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild]()

Datasets – Sports-1M

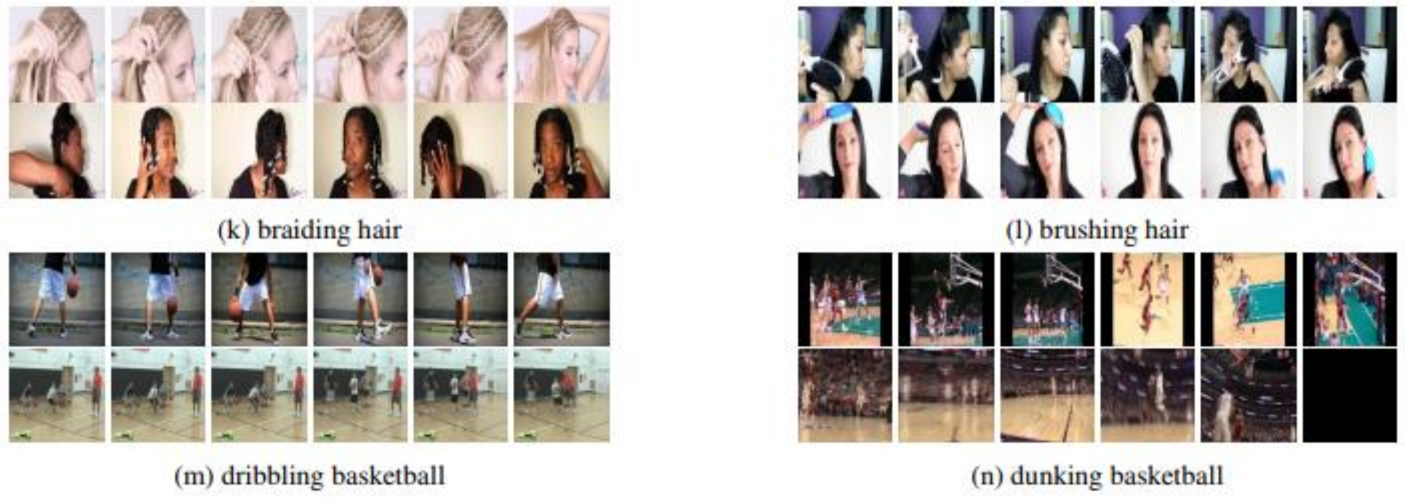
- 1,133,158 videos from YouTube
- 487 sport categories
- Automatically labelled by analyzing the text metadata



Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks., CVPR

Datasets – Kinetics

- 400 human action classes
- 240k training videos
- Manual annotations
- Person Actions (singular), e.g. drawing, drinking, laughing, punching; Person-Person Actions, e.g. hugging, kissing, shaking hands; and, Person-Object Actions, e.g. opening presents, mowing lawn, washing dishes



Kay et al (2017), The Kinetics Human Action Video Dataset

Challenges in video classification

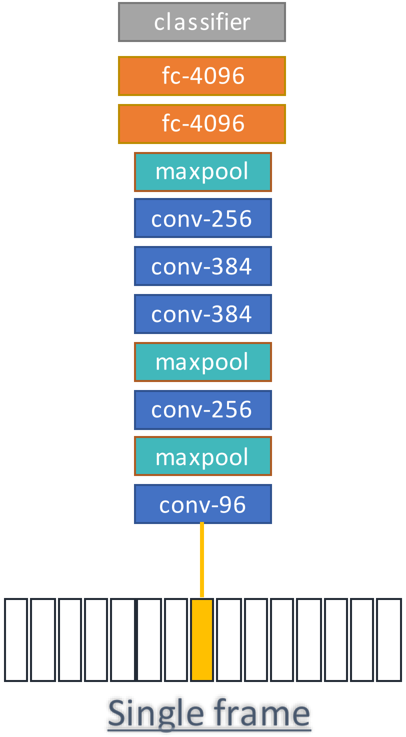
- Computationally expensive
  - Number of frames >> number of images
- Lower image quality
  - Resolution, motion blur, occlusion
- Weak labels
  - Video-level labels

Video as a sequence of images

- Let’s use CNN (AlexNet) as a backbone
- **Question 1**: How do we integrate predictions from individual frames of a video?

- Split each video into K x N-frame clips
- Average their predictions over K clips

- Single frame architecture predicts the category of middle frame for each clip

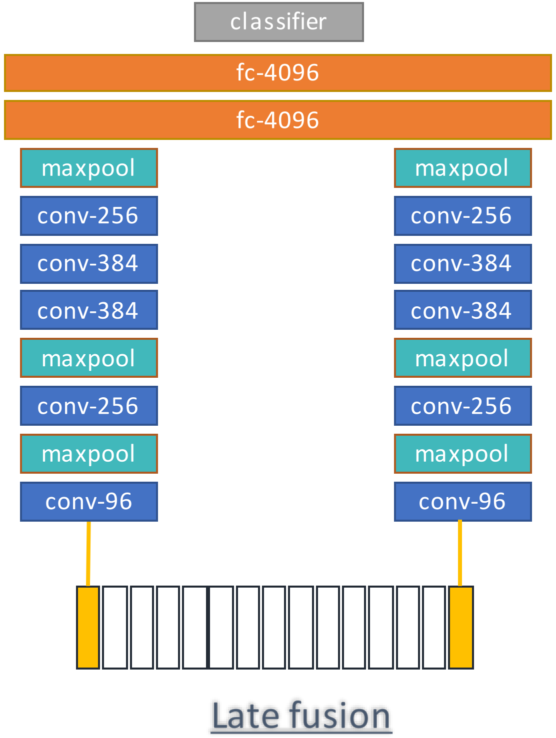


[Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks, CVPR]

Late fusion

Early fusion architecture

- Uses two frames as input per clip
- Parameters shared across two towers
- Merges their features after the last convolutional layer
- Doubles number of filters in the first fully connected layer
  (e.g. 6x6x**256**x4096 to
6x6x**512**x4096)
- Compares high level features from two frames

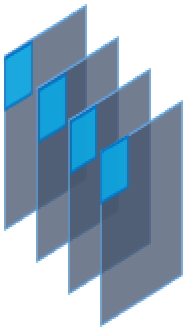


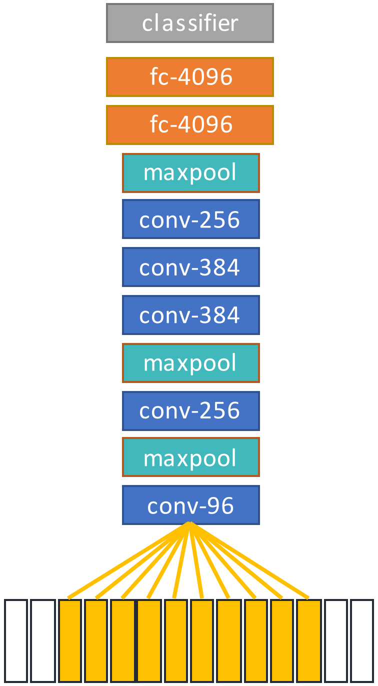
[Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks, CVPR]()

Early fusion

Early fusion architecture

- Uses 10 frames as input per clip
- Concatenates them at pixel level
  (HxWx**3**x**10** to HxWx**30**x1)
- 10 times more filters in the first convolution layer, *e.g.* 11x11x**3**x96 to 11x11x**30**x96
- Compares low level features from ten frames
- Can detect only local motion





Early fusion

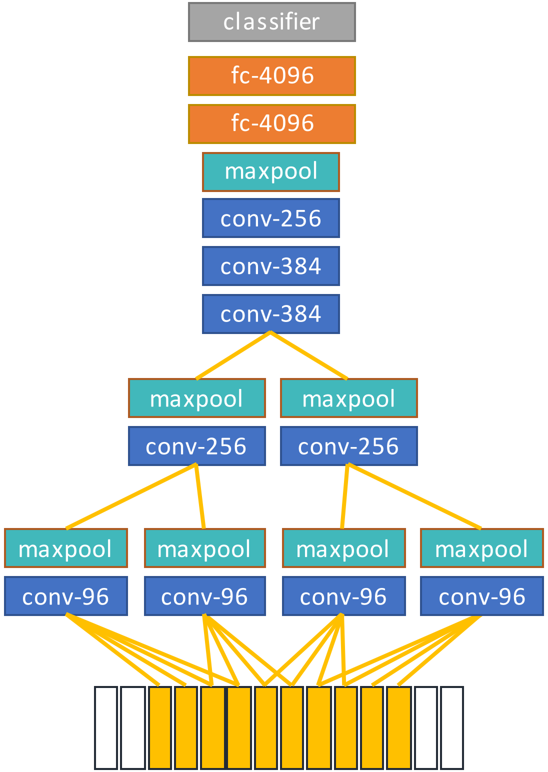
[Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks, CVPR](#)

Slow fusion

Slow fusion architecture

- Mix of early and late fusion
- Uses 10 frames as input per clip
- Extends connectivity of convolutional layers in time in addition to spatial convolutions

$$H \times W \times F_{in} \times F_{out}$$
$$\rightarrow \mathbf{T} \times H \times W \times F_{in} \times F_{out}$$

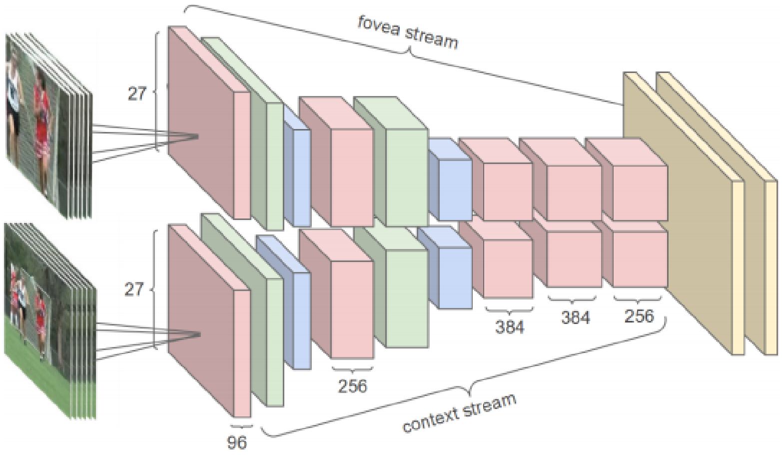


Slow fusion

Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks, CVPR

Multi-resolution: fovea and context

- **Question 2**: How can we efficiently train over millions of frames?
- Uses two networks that focus on
    - A smaller image region, central patch (fovea)
    - Whole frame on half res (context)
    - Two inputs of (H/2)x(W/2)x3 instead of HxWx3
- Wikipedia: The fovea centralis is a small, central pit composed of closely packed cones in the eye.



[Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks., CVPR](#)

Results

- Dataset: Sports 1M, 1 million YouTube videos annotated with 487 classes
- Trained on ~50M frames

- Clip-level prediction
  - 0.5 second length sequences from videos
  - labels are noisy

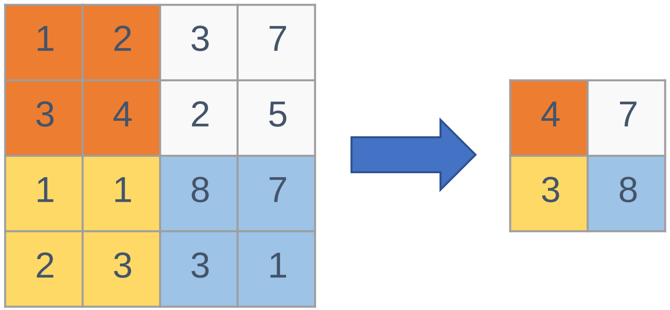
- Video-level prediction
  - randomly sample 20 clips
  - feed each clip individually to the network
  - average the scores

Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
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

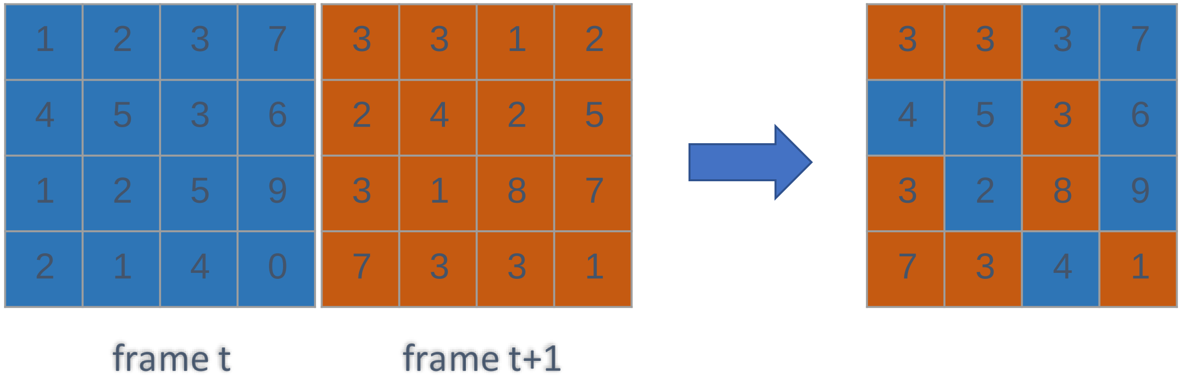
Karpathy et al (2014), Large-scale Video Classification with Convolutional Neural Networks., CVPR

Temporal max pooling

Temporal pooling

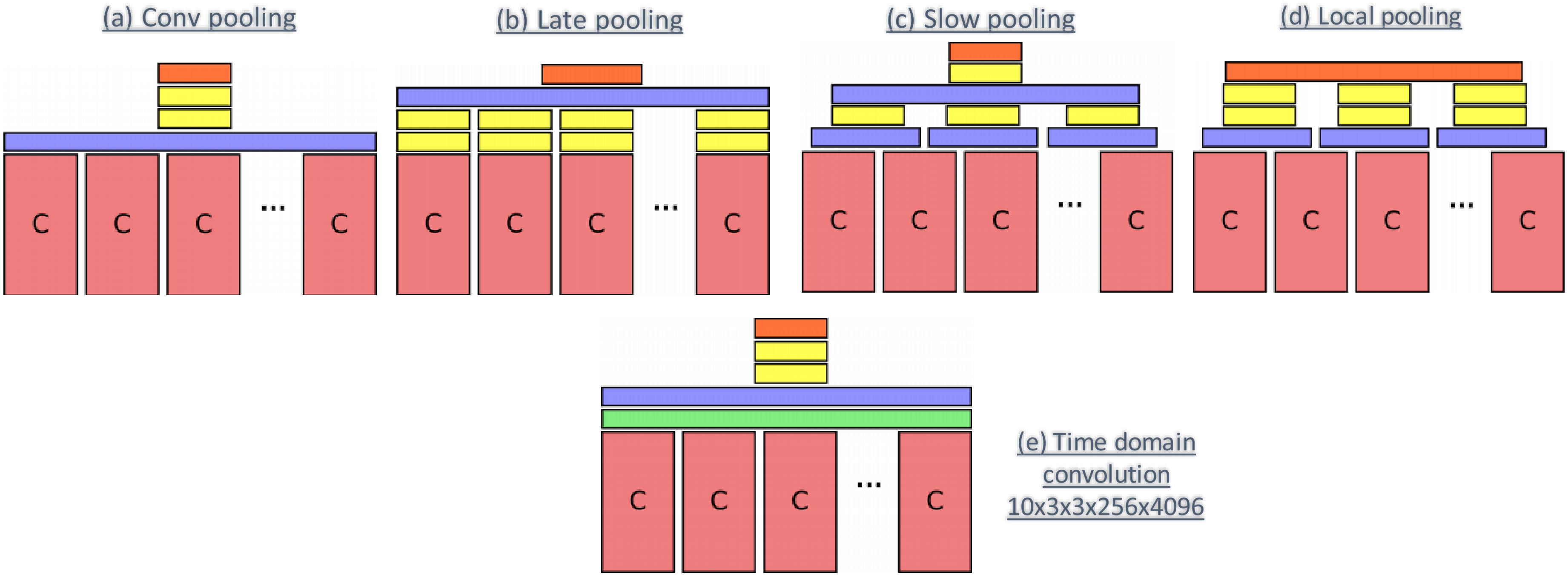


Spatial pooling



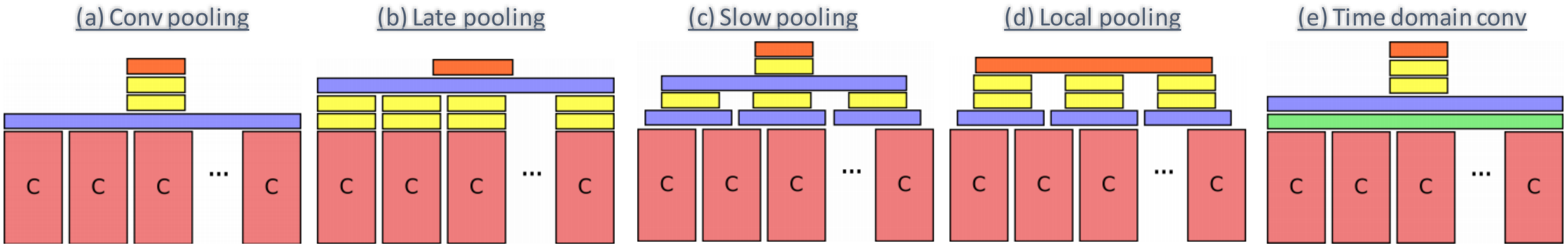
Temporal feature pooling

- Claim: Previous work uses short clips (0.5 sec). An accurate prediction requires a global view on videos.



[Yue-Hei Ng et al. (2015), Beyond short snippets: Deep networks for video classification. CVPR.](#)

Results: temporal feature pooling



- 120 frame AlexNet model
- Max temporal pooling over last conv layer performs best
- Preserving spatial information during temporal pooling is important
- Time-domain convolution is not effective in learning temporal relations

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

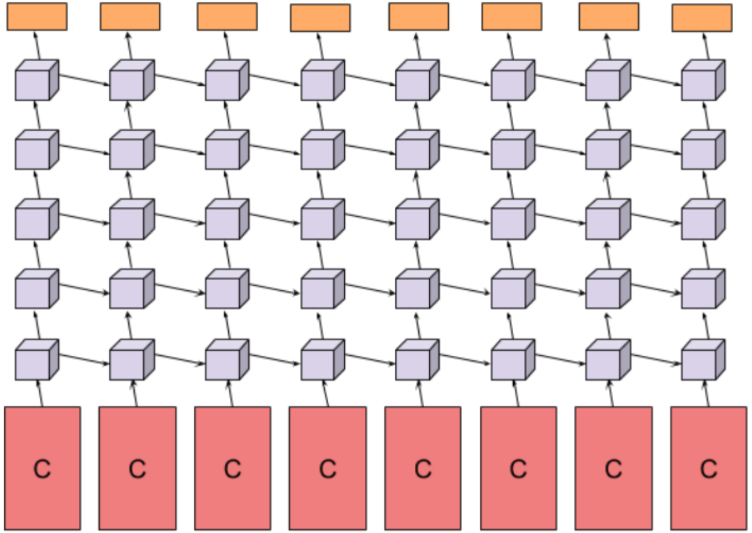
Yue-Hei Ng et al. (2015), Beyond short snippets: Deep networks for video classification. CVPR.

CNN + LSTM

- Observation: During temporal max pooling, temporal order is lost
- Hypothesis: LSTM encodes temporal relations better, thus LSTM on CNN features should be a better model

Dataset: Sports 1M
Evaluated with 2 network architectures
- AlexNet and GoogleNet

GoogleNet outperforms AlexNet
Conv pooling outperforms LSTM



Method	Network	Frames	Video Hit@1	Video Hit@5
Conv pooling	AlexNet	120	71.1	89.3
Conv pooling	GoogleNet	120	72.3	90.8
LSTM	AlexNet	30	62.7	83.6
LSTM	GoogleNet	30	72.1	90.4

Yue-Hei Ng et al. (2015), Beyond short snippets: Deep networks for video classification. CVPR.

Motion

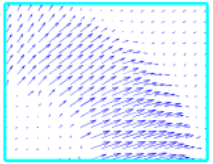
Even “impoverished” motion data can evoke a strong percept

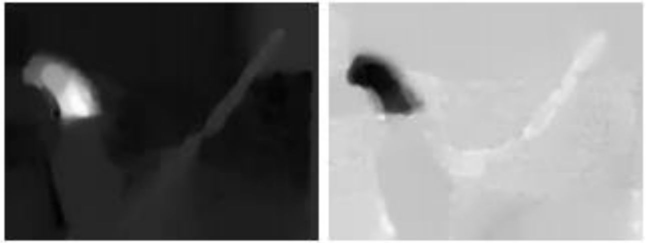
[G. Johansson (1973), Visual Perception of Biological Motion and a Model For Its Analysis](#)

Optical flow

- So far, video = sequence of frames captured over time
- Alternative, video = appearance + motion
- Optical flow: displacement of a pixel over time
  - $I(x, y, t + \Delta_t) = I(x + \Delta_x, y + \Delta_y, t)$
  - Two channel input: $\Delta_x(x, y, t)$, $\Delta_y(x, y, t)$

$I(x, y, t)$ $I(x, y, t + \Delta_t)$





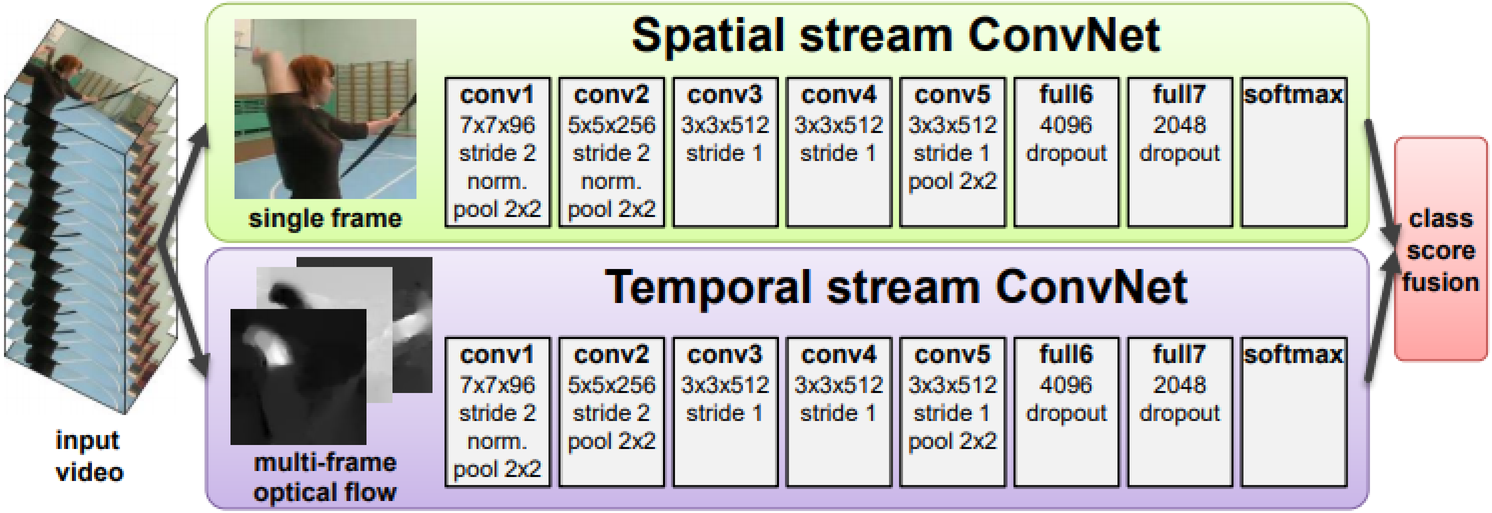
Δ_x Δ_y

Simonyan and Zisserman (2014), Two-Stream Convolutional Networks, NIPS.

Two stream network

Previous work: It can be difficult to learn the concept of motion implicitly
Proposal: This work separates motion from static appearance

- Motion: external + camera → mean subtraction to compensate camera motion
- Stacks 10 optical flow frames



Slide Credit

Simonyan and Zisserman (2014), Two-Stream Convolutional Networks, NIPS.

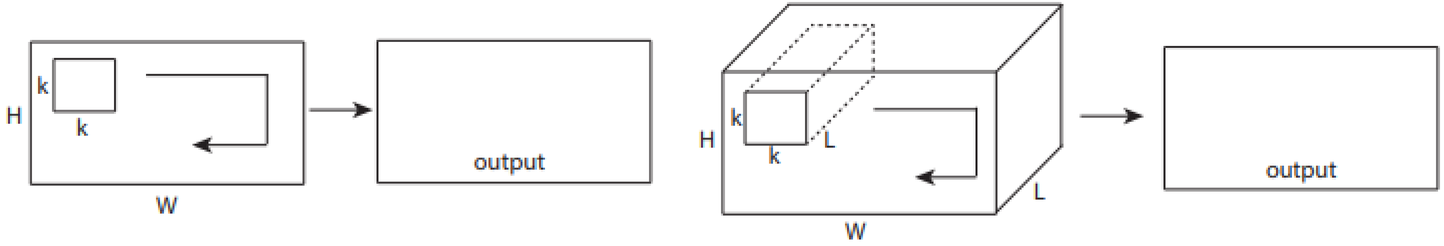
Results: Two stream network

- Base model is VGG-M
- Datasets: UCF101 and HMDB51
- Spatial ConvNet is pre-trained on ImageNet
- Temporal ConvNet is trained from scratch
- Temporal and spatial recognition streams are complementary

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

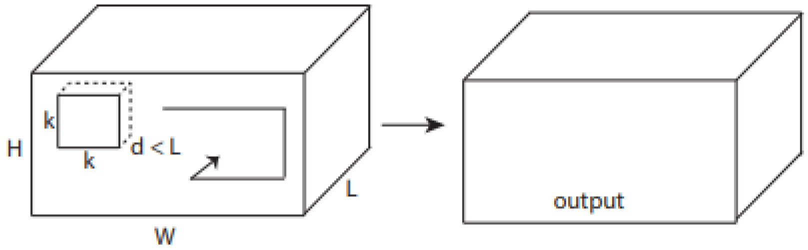
3D convolutions

Problem: Temporal ordering is lost in 2D convolutions
Idea: A natural way to deal with 3D data is 3D convolutions



(a) 2D convolution on an image

(b) 2D convolution on a video



(c) 3D convolution on a video

[Ji et al (2013), 3D Convolutional Neural Networks for Human Action Recognition, TPAMI.]()
[Tran et al (2015), Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV.]()

3D convolutional networks

- 3x3x3 convolution kernels with stride 1
- 2x2x2 pooling kernels (except *pool1* 1x2x2)
- Works on 16 frame-length clips
- Trained from scratch on Sports-1M dataset

Perfo
Conv

Method	Number of Nets	Clip hit@1	Video hit@1	Video hit@5
DeepVideo's Single-Frame + Multires [18]	3 nets	42.4	60.0	78.5
DeepVideo's Slow Fusion [18]	1 net	41.9	60.9	80.2
Convolution pooling on 120-frame clips [29]	3 net	**70.8***	**72.4**	**90.8**
C3D (trained from scratch)	1 net	44.9	60.0	84.4
C3D (fine-tuned from I380K pre-trained model)	1 net	46.1	61.1	85.2

Tran et al (2015), Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV.

Mixed 3D-2D convolutional networks

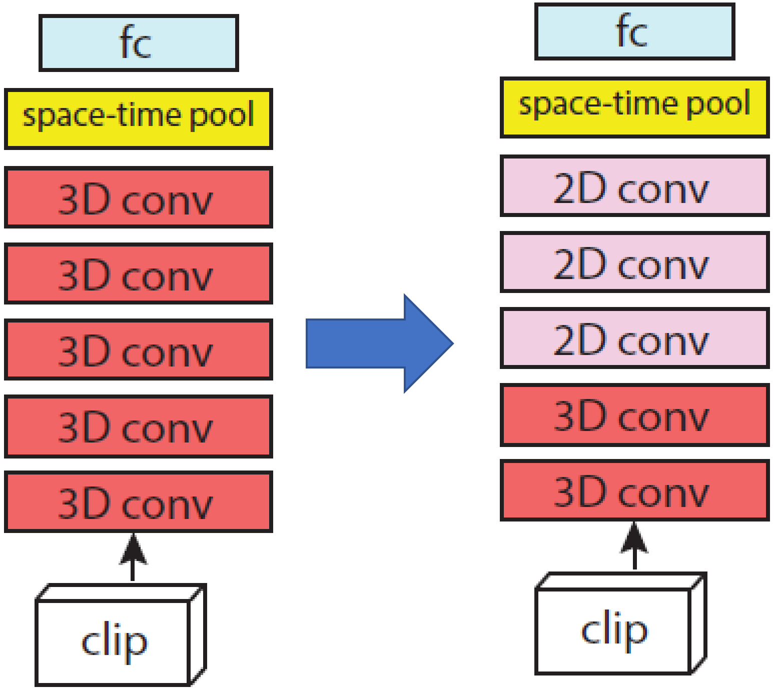
Observation

- 3D convs have 3x more parameters than 2D convs (to learn)

Hypothesis

- Motion is a low/mid-level concept so it should be implemented in early layers

Network	# parameters	Video Hit@1
2D	11.4M	59.5
3D(1x)+2D	11.4M	61.8
3D(2x)+2D	11.7M	62.5
3D(3x)+2D	12.7M	62.9
3D(4x)+2D	16.9M	62.5
3D(all)	33.4M	61.8



Tran et al (2018), A Closer Look at Spatiotemporal Convolutions for Action Recognition, CVPR.

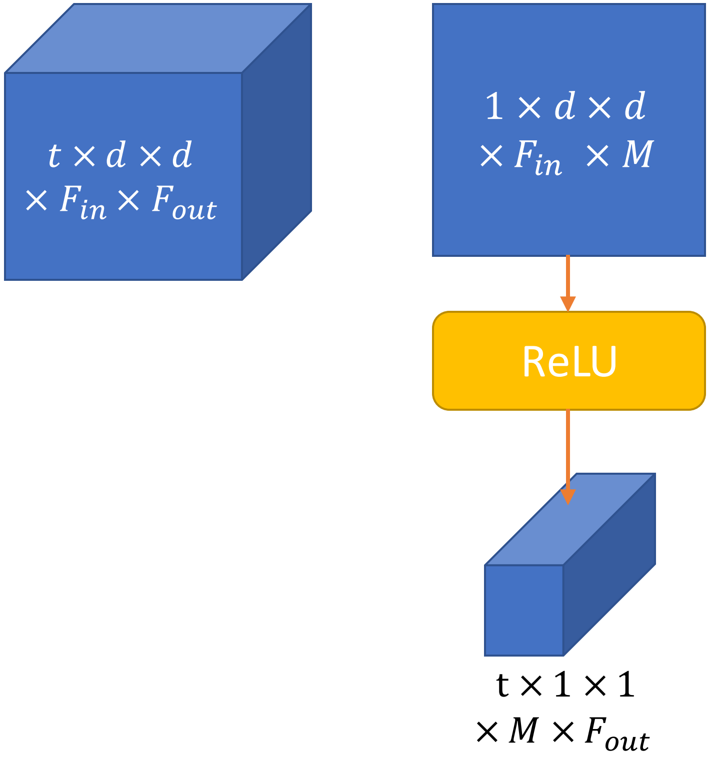
(2+1)D convolutions

Observation: 3D convs can be factorized into 2D+1D convolutions

2D spatial conv + ReLU + 1D temporal conv

$$\mathrm{d} \times d \times F_{in} \times F_{out}$$
$$\boldsymbol{t} \times H \times W \times F_{in} \times F_{out}$$

Network	# parameters	Video Hit@1
2D	11.4M	59.5
3D(3x)+2D	12.7M	62.9
3D(all)	33.4M	61.8
(2+1)D	33.3M	64.8



Tran et al (2018), A Closer Look at Spatiotemporal Convolutions for Action Recognition, CVPR.

Summary

Improvements in video classification

- Larger datasets (Sports-1M, Kinetics)
- Motion information (optical flow, temporal convolutions/pooling, 3D)
- Better architectures (ResNets)

Recommended reading

- Yue-Hei Ng et al. (2015), Beyond short snippets: Deep networks for video classification. CVPR.

Additional reading

- Tran et al (2018), A Closer Look at Spatiotemporal Convolutions for Action Recognition, CVPR.