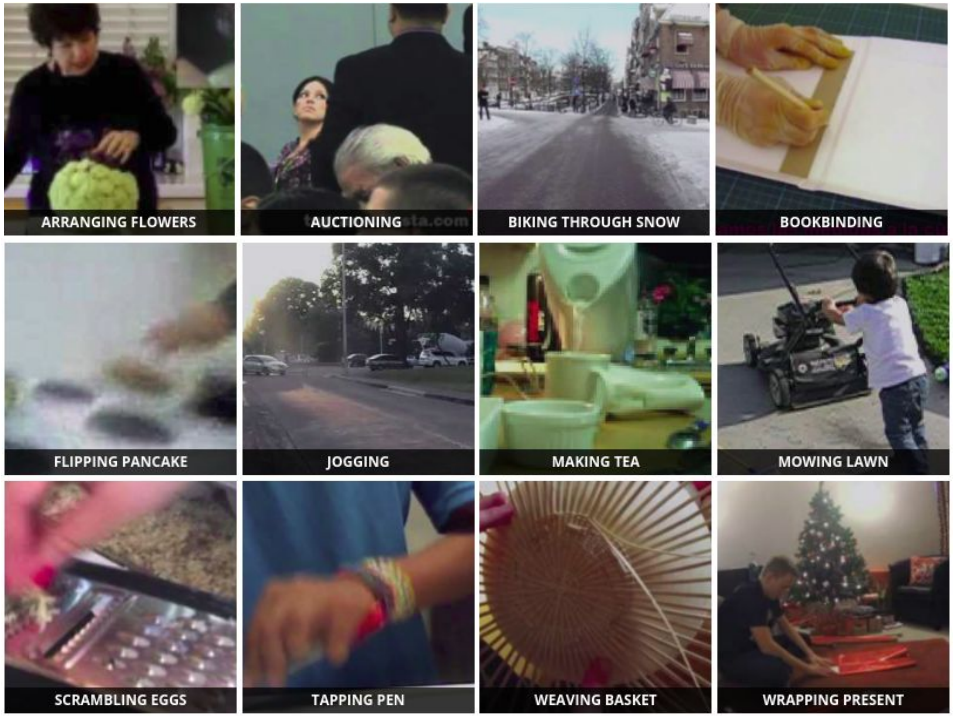
Deep Neural Networks for Action Recognition in Videos

Ondřej Bíža
Showmax Lab
Faculty of Information Technology,
Czech Technical University
Prague, Czech Republic

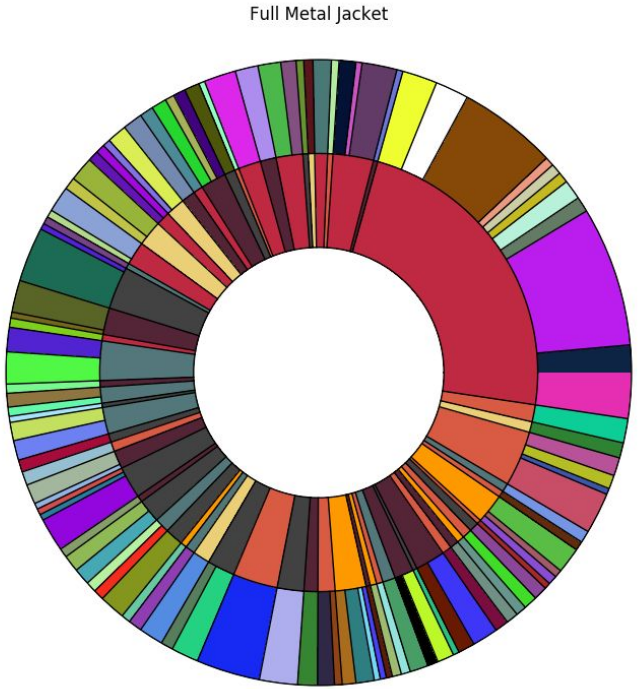
Human-Focused Action Recognition



source: Kinetics dataset

- videos featuring people performing causal actions
- teach a Machine Learning model to recognize what is happening in the videos
- applications: intelligent video surveillance, human-computer interaction, video browsing and recommendation

Scene detection and feature extraction

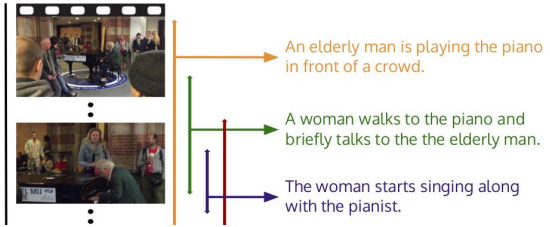


Tasks

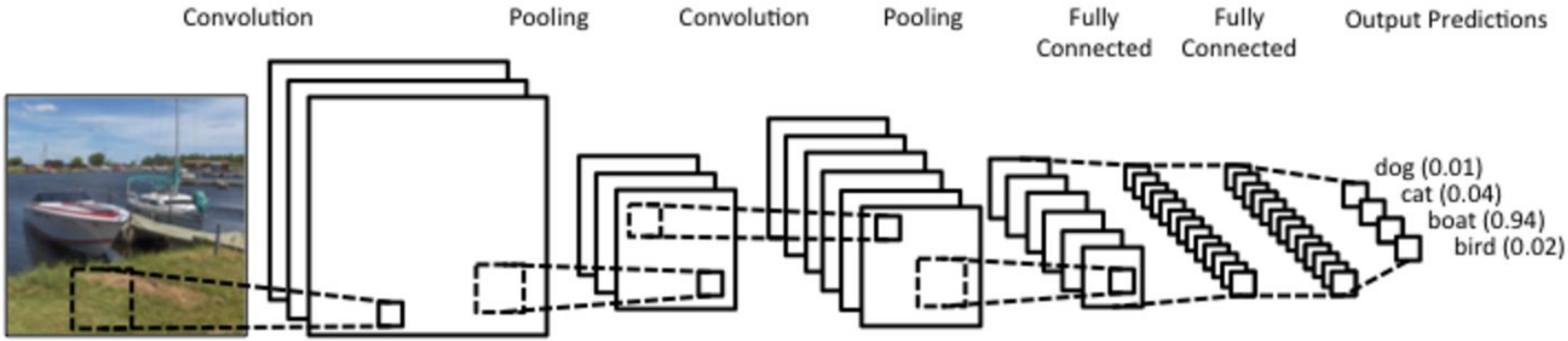
Classification

Localization

Captioning



Convolutional Neural Network (ConvNet)



source

- local receptive fields model local structures
- low number of weights due to weight sharing -> lower tendency to overfit
- invariance to translation

Filter Visualization

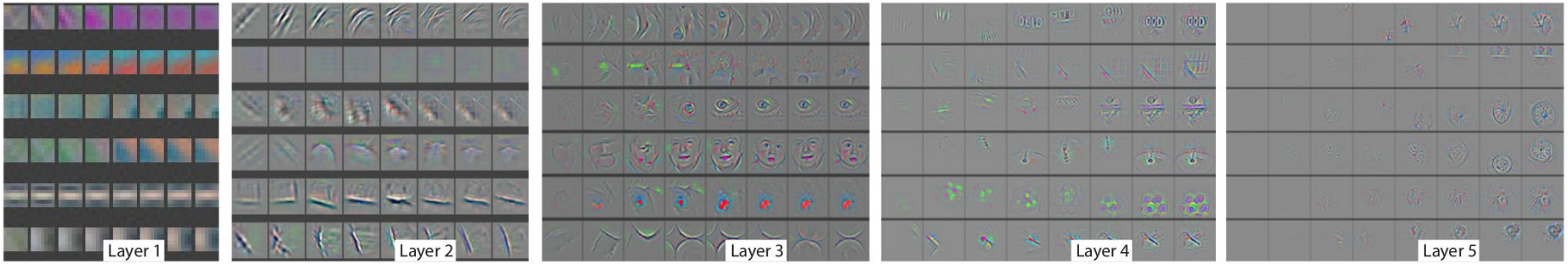
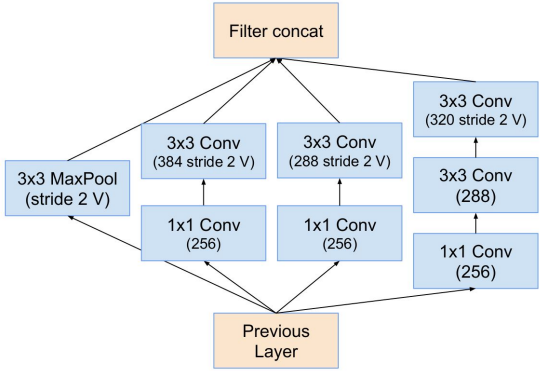
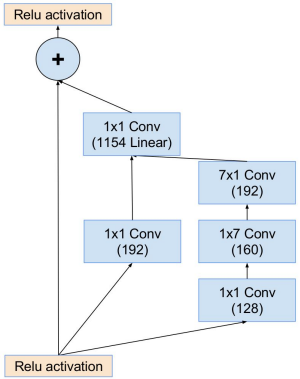


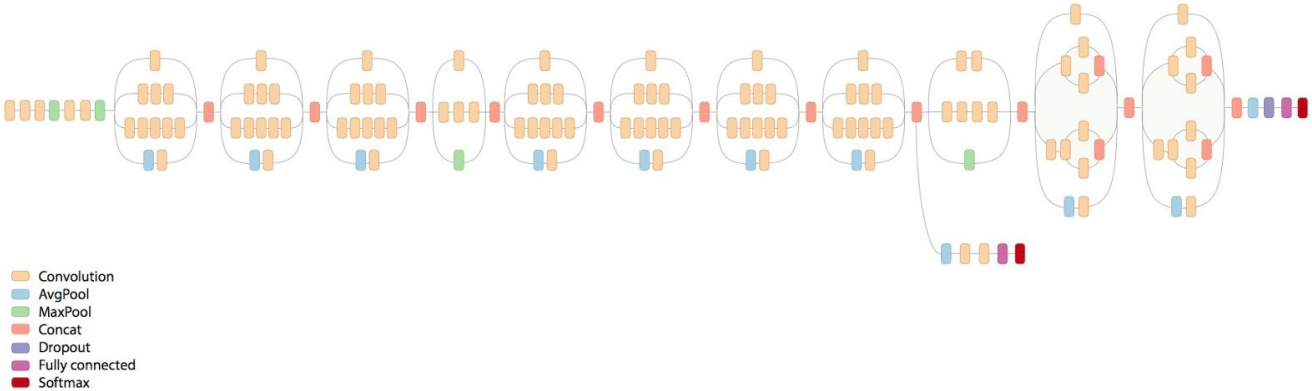
Figure 4. Evolution of a randomly chosen subset of model features through training. Each layer's features are displayed in a different block. Within each block, we show a randomly chosen subset of features at epochs [1,2,5,10,20,30,40,64]. The visualization shows the strongest activation (across all training examples) for a given feature map, projected down to pixel space using our deconvnet approach. Color contrast is artificially enhanced and the figure is best viewed in electronic form.

Visualizing and Understanding Convolutional Networks - M.D.Zeiler et al. (2013)

Inceptions







Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning - C. Szegedy et al. (2016)

Residual Network (ResNet)

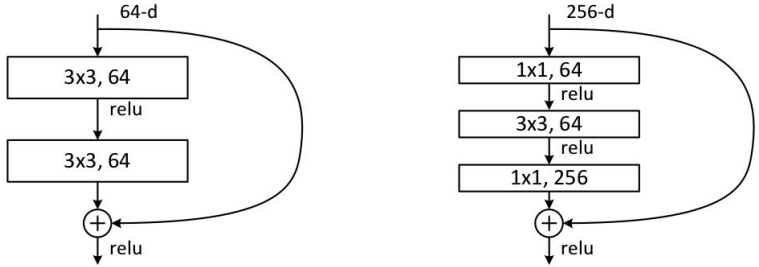
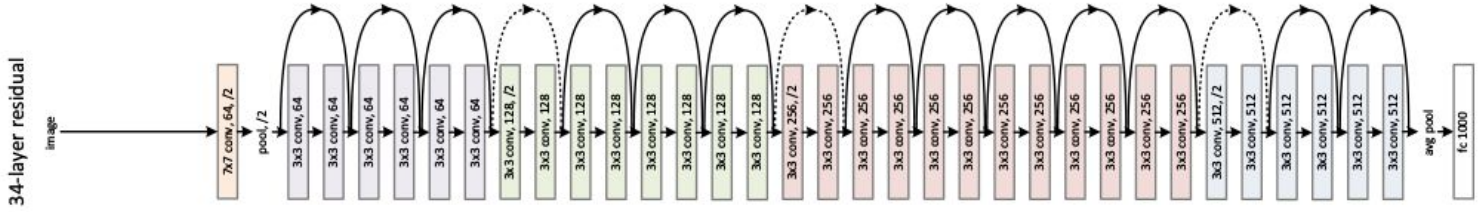


Figure 5. A deeper residual function $\mathcal{F}$ for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a "bottleneck" building block for ResNet-50/101/152.



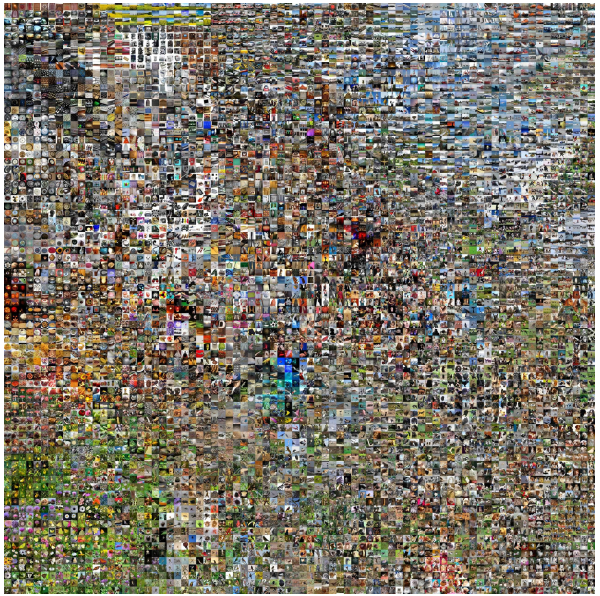
Deep Residual Learning for Image Recognition - K. He et al. (2015)

The dataset

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

ImageNet 2012 / ImageNet-1k / ILSVRC 2012

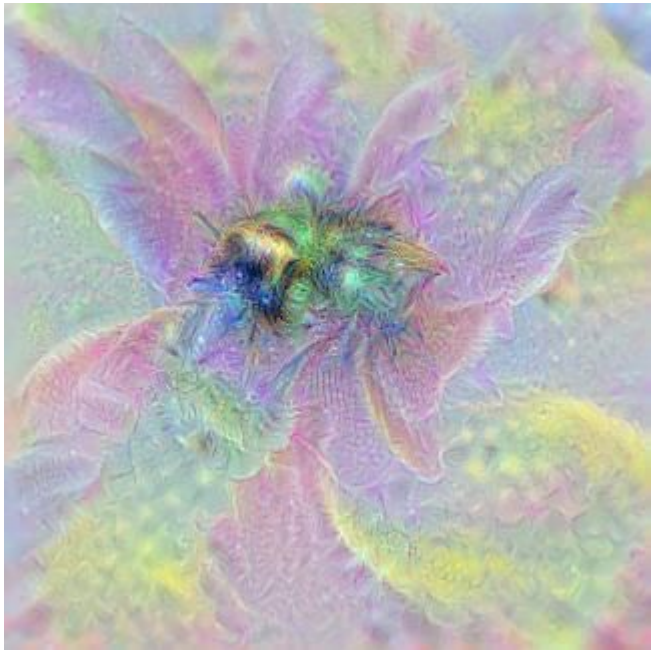
- 1000 object classes
- 1.2M training images
- 100k testing images



http://cs.stanford.edu/people/karpathy/cnnembed/

The World According to Inception-v1

Bear



Bee

source

The World According to Inception-v1

Saxophone

source

Datasets

- HMDB-51, UCF-50 and UCF-101
- DeepMind’s Kinetics

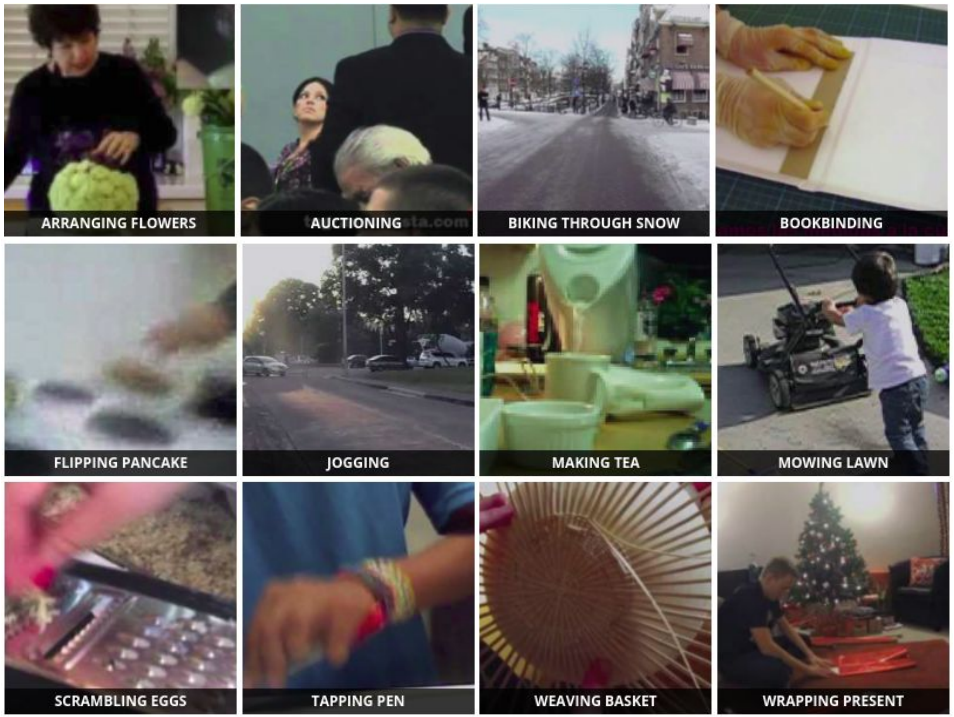
UCF-101 (2012)

~ 13000 sequences
101 classes

source: UCF-101 dataset

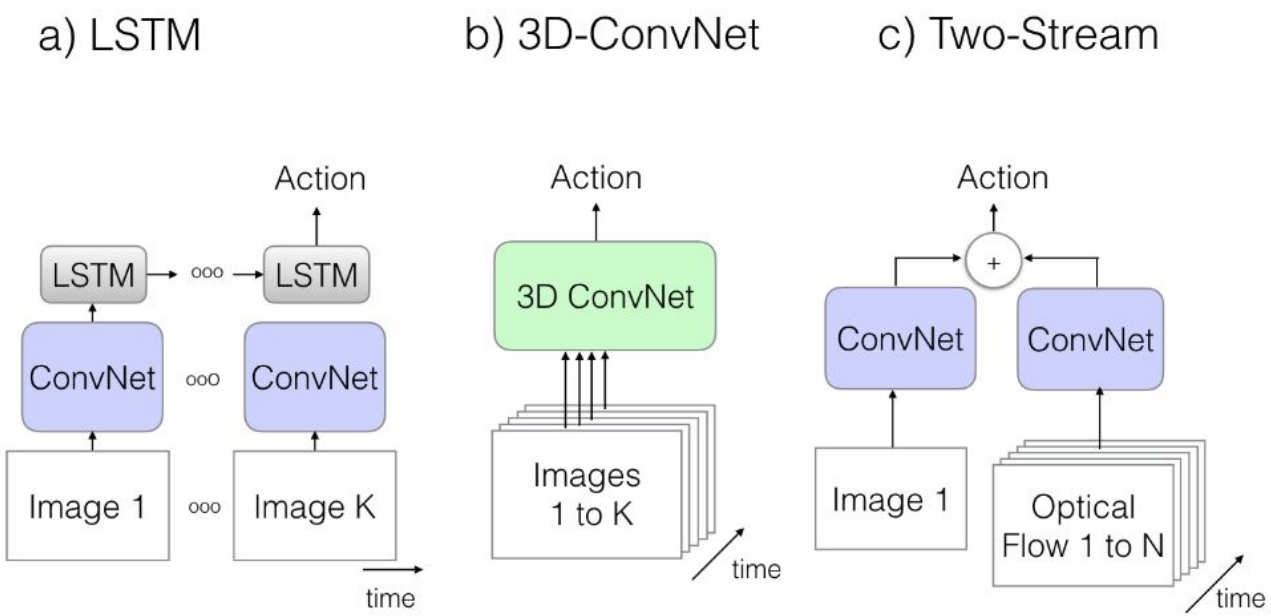
DeepMind's Kinetics (2017)

~ **300000** sequences
400 classes



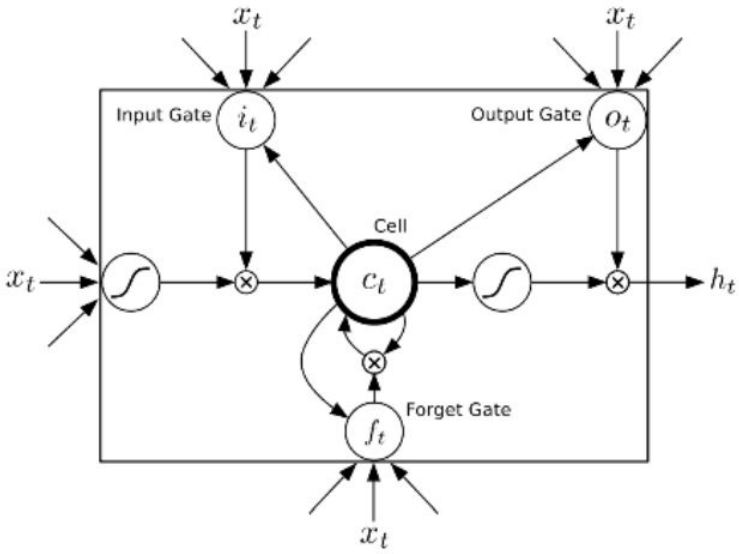
source: Kinetics dataset

Three Approaches to Modelling Video

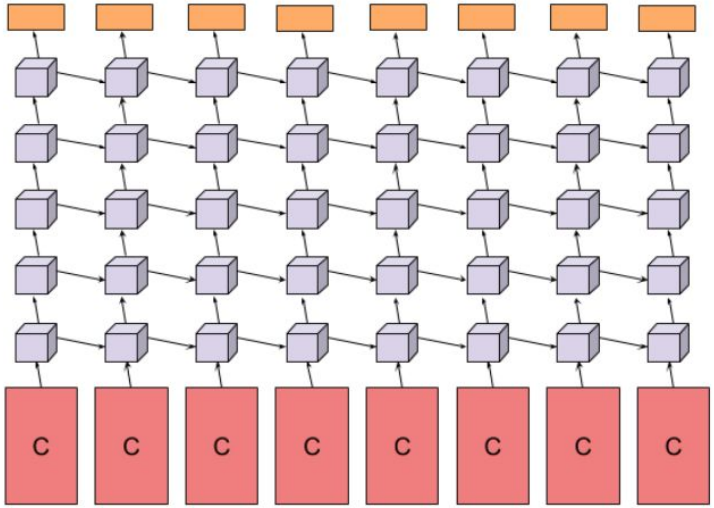


Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset - J.Carreira et al. (2017)

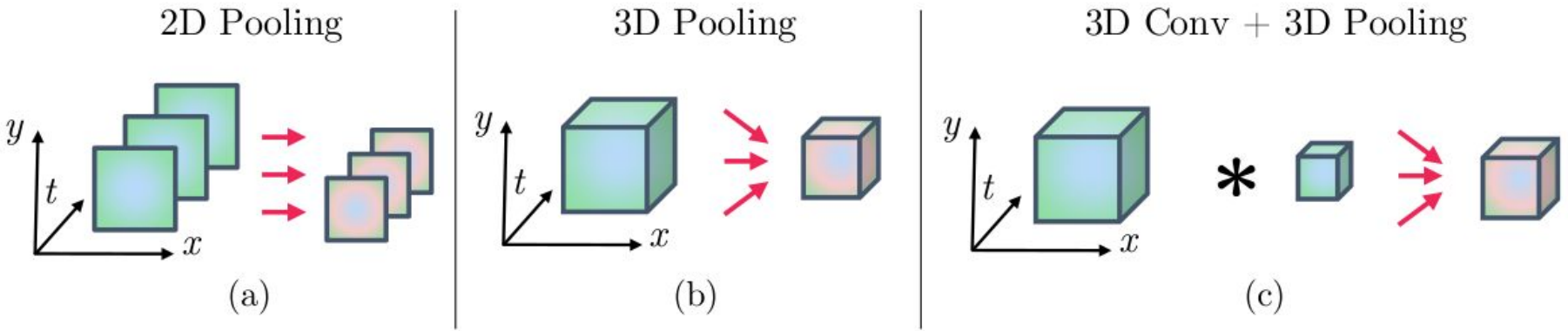
RNN + ConvNet



Beyond Short Snippets: Deep Networks for Video Classification - J.Y. Ng et al. (2015)

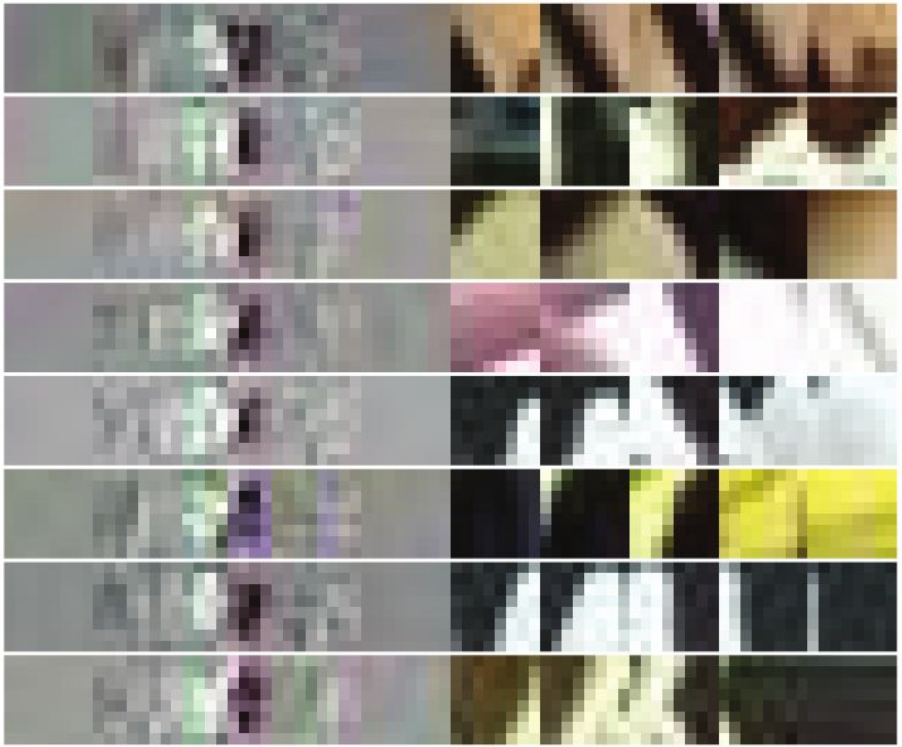


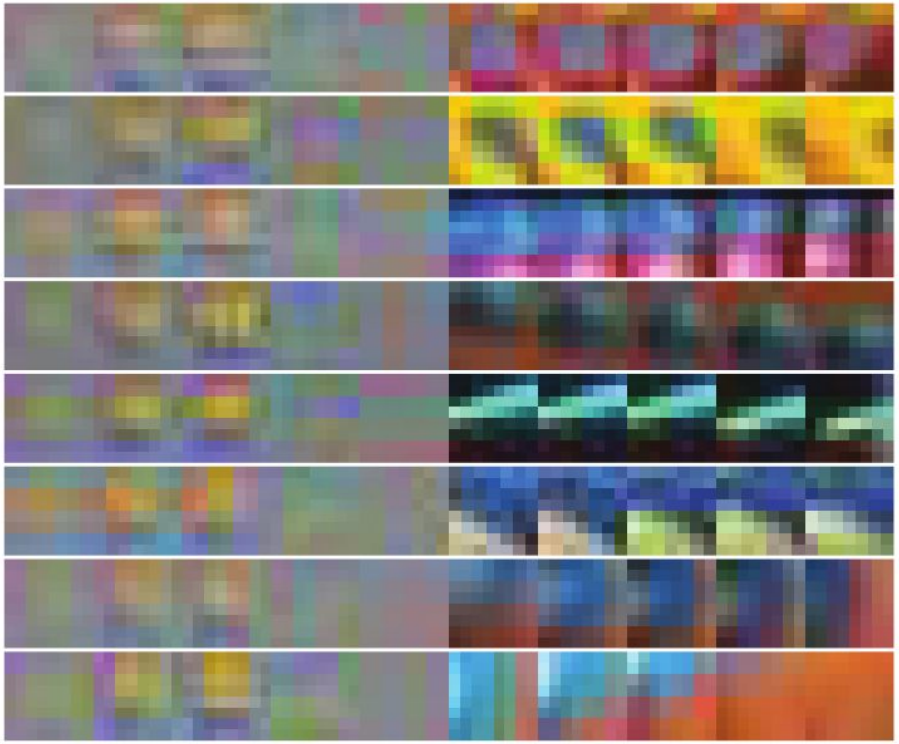
3D-ConvNet / C3D



Convolutional Two-Stream Network Fusion for Video Action Recognition - C. Feichtenhofer et al. (2016)

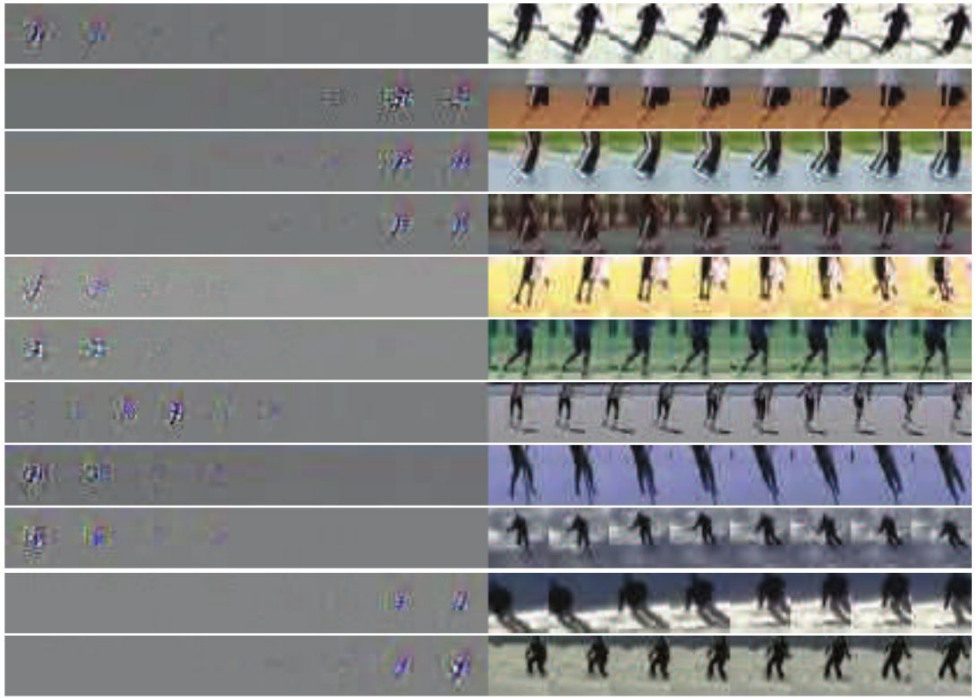
What do the 3D filters learn?





Learning Spatiotemporal Features with 3D Convolutional Networks - D. Tran et al. (2014)

What do the 3D filters learn?



Learning Spatiotemporal Features with 3D Convolutional Networks - D. Tran et al. (2014)

What do the 3D filters learn?

Learning Spatiotemporal Features with 3D Convolutional Networks - D. Tran et al. (2014)

Optical Flow

(a) First frame (b) Second frame (c) Optical flow field

https://www.semanticscholar.org/paper/A-Duality-Based-Approach-for-Realtime-TV-L1-Optica-Zach-Pock/0f6bbe9afab5fd61f36de5461e9e6a30ca462c7c

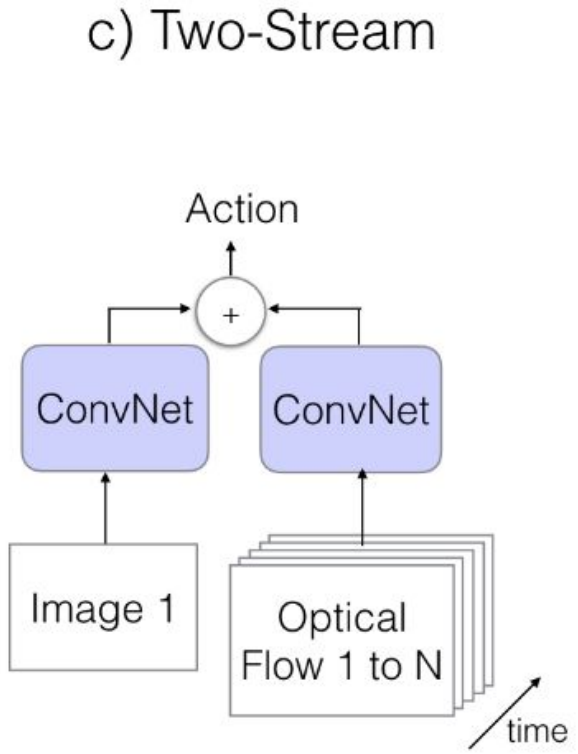
Optical Flow Example Video

Two-Stream

Loosely inspired by neuroscience.

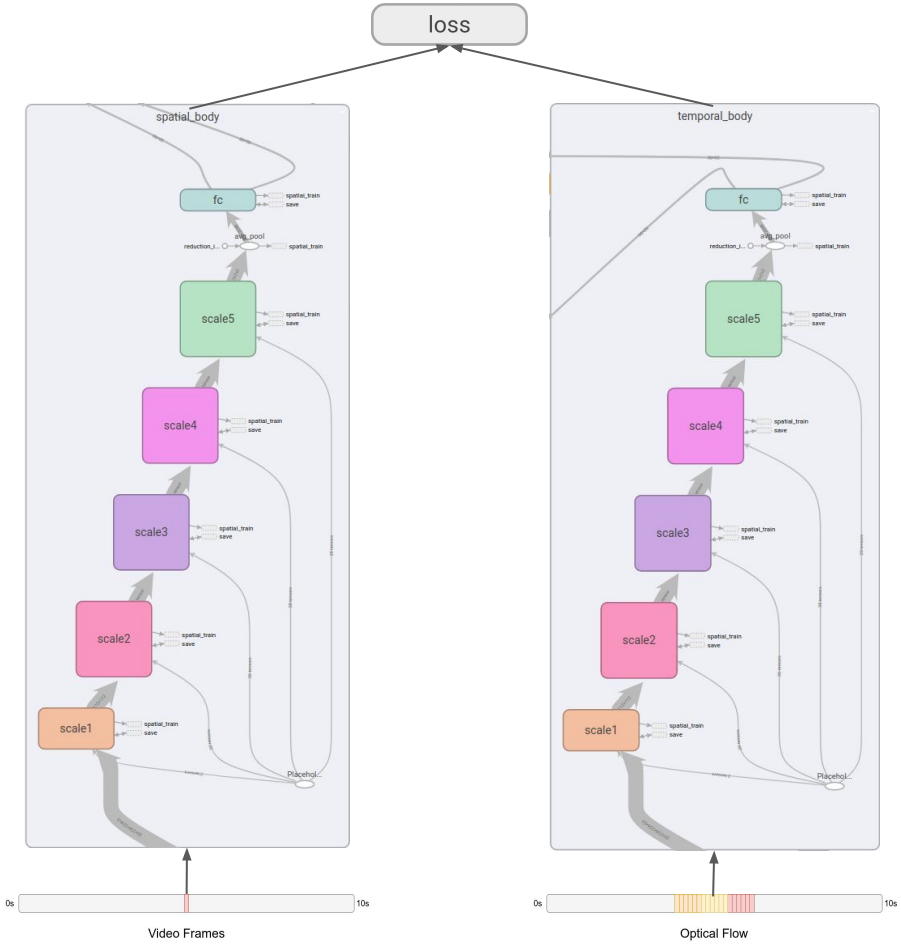
- **ventral stream**: identification of objects
- **dorsal stream**: “mediates the required sensorimotor transformations for visually guided actions directed at such objects”

Reference: Separate visual pathways for perception and action - M.A. Goodale et al. (1992).



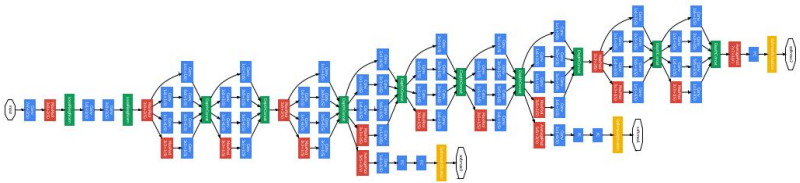
Two-Stream

- the temporal stream needs high-quality optical flow to recognize actions, this poses two problems
  - high-quality optical flow is very **expensive to compute** (e.g. TV-L1 algorithm runs at 14 fps on a high-end GPU)
  - videos with significant **camera motion** confuse the ConvNet
- overall, the Two Stream architecture is quite inconvenient



Architectures Comparison - Methodology

- use Inception-v1 (also called GoogLeNet) as the 2D ConvNet
- optical flow is computed using TV-L1 algorithm
- basic data augmentation during training



Inception-v1 / GoogLeNet

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

Comparison

Architecture	UCF-101 accuracy (%)	HMDB-51 accuracy (%)
RNN + ConvNet	81.0	36.0
3D-ConvNet	49.2	24.3
Two-Stream	91.2	58.3

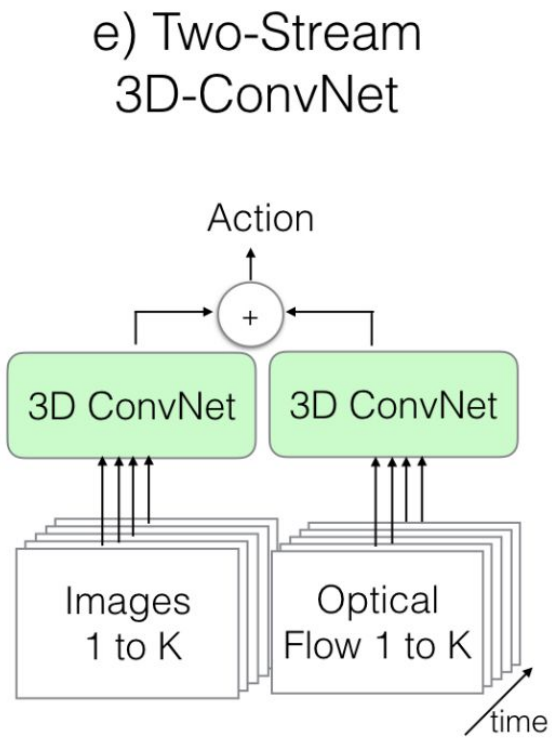
Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset - J.Carreira et al. (2017)

Comparison - more training data

Architecture	**UCF-101 accuracy (%)**	**HMDB-51 accuracy (%)**
RNN + ConvNet	82.1 (81)	46.4 (36)
3D-ConvNet	79.9 (51.6)	49.4 (24.3)
Two-Stream	91.5 (91.2)	58.7 (58.3)

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset - J.Carreira et al. (2017)

State-of-the-art: I3D



Without pretraining (accuracy %):

UCF-101	-	93.4
HMDB-51	-	66.4
Kinetics	-	**74.2**

With pretraining on Kinetics (accuracy %):

UCF-101	-	98
HMDB-51	-	80.7

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset - J.Carreira et al. (2017)

I3D Ablation Analysis

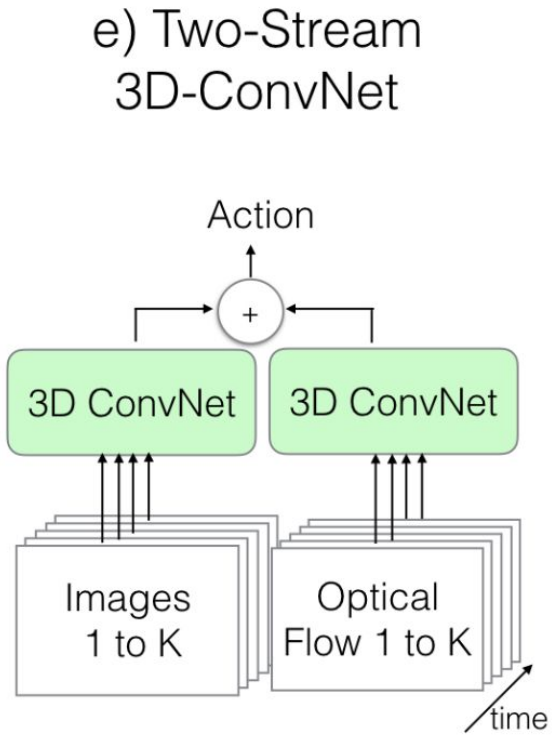
Kinetics dataset

I3D: 74.2%

I3D without ImageNet pretraining: 71.6% (-2.6%)

appearance stream only: 71.1% (-3.1%)

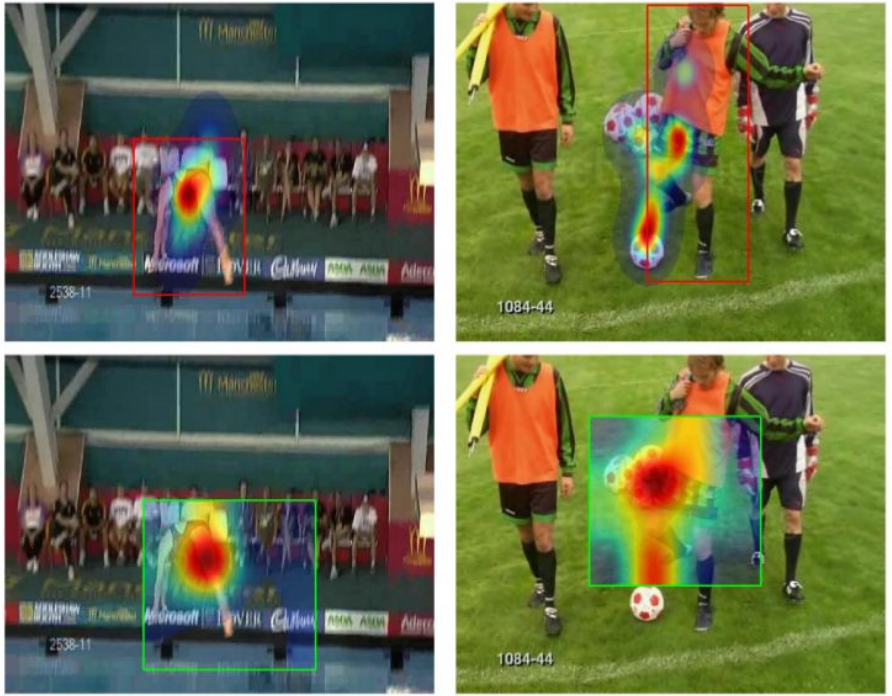
motion stream only: 63.4% (-10.8%)



Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset - J.Carreira et al. (2017)

Attention

Action Recognition using Visual Attention - S. Sharma et al. (2015)



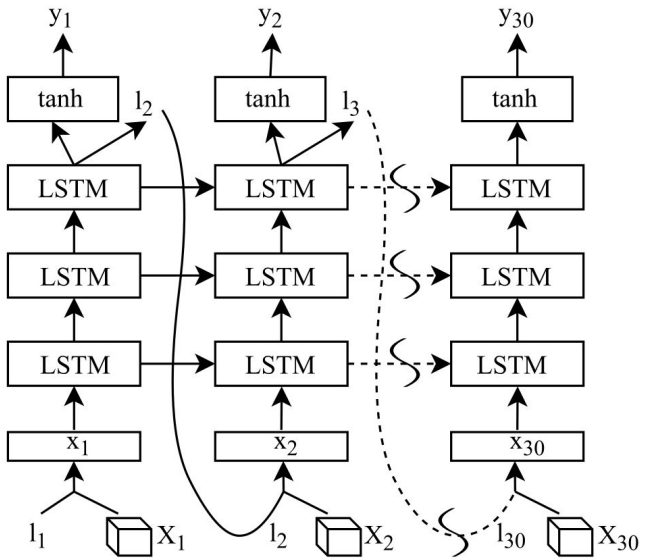
Action is in the Eye of the Beholder: Eye-gaze Driven Model for Spatio-Temporal Action Localization - N. Shapovalova et al. (2013)

Learning to Attend

Explicit Training

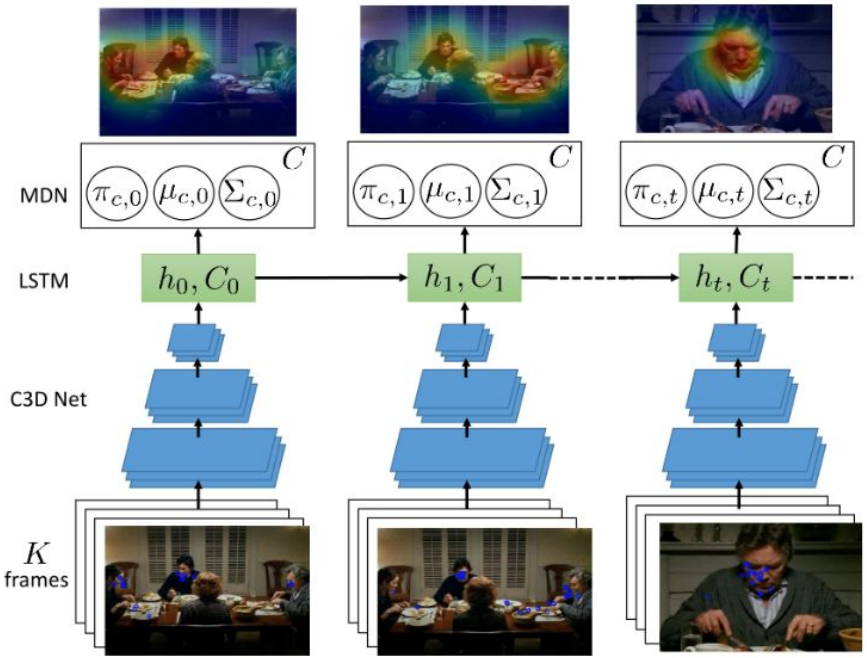
http://www.eyegaze.com/4-eye-tracking-technology-applications-you-may-not-know/

Implicit Training



Action Recognition using Visual Attention - S. Sharma et al. (2015)

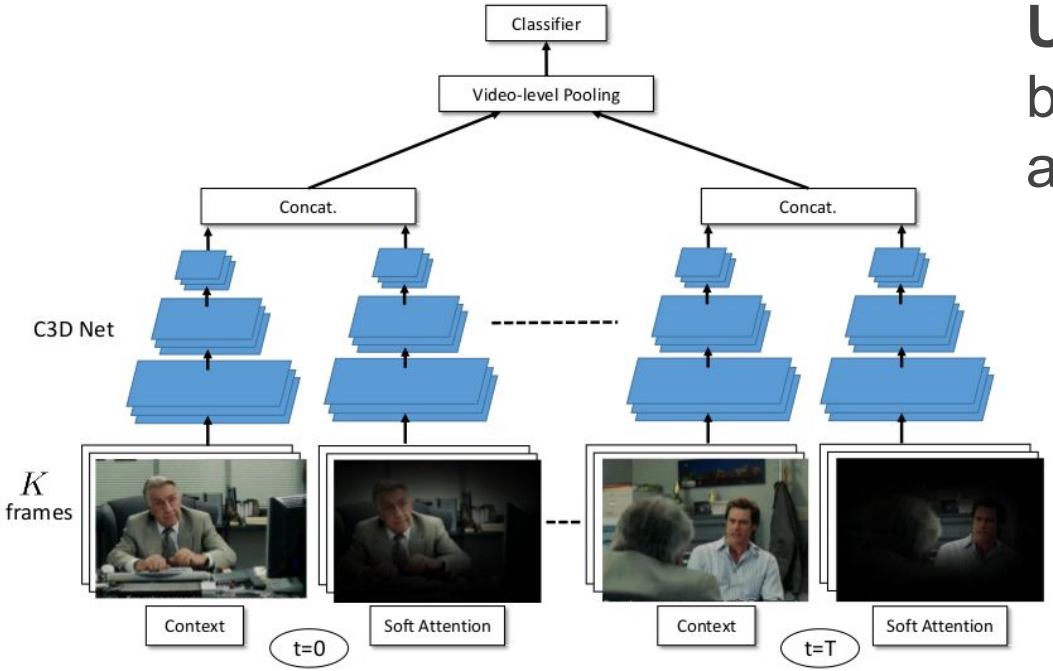
Explicit Attention Training



- 3D ConvNet models short snippets of videos
- Recurrent Neural Network (LSTM) models long-term dynamics
- The model is trained to predict human fixations for each frame

Recurrent Mixture Density Network for Spatiotemporal Visual Attention - L. Bazzani et al. (2016)

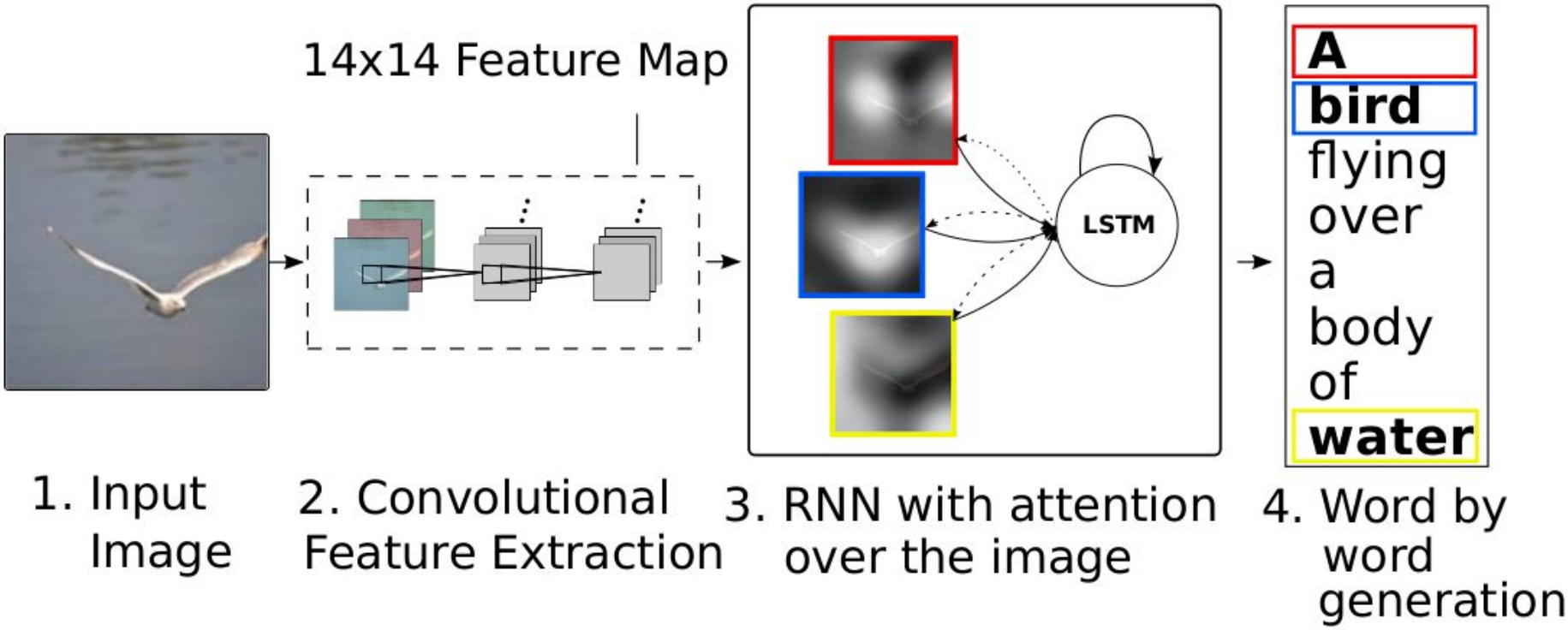
Explicit Attention Training



UCF-101 dataset
baseline: 80.4%
attention: 82.8% (+2.4%)

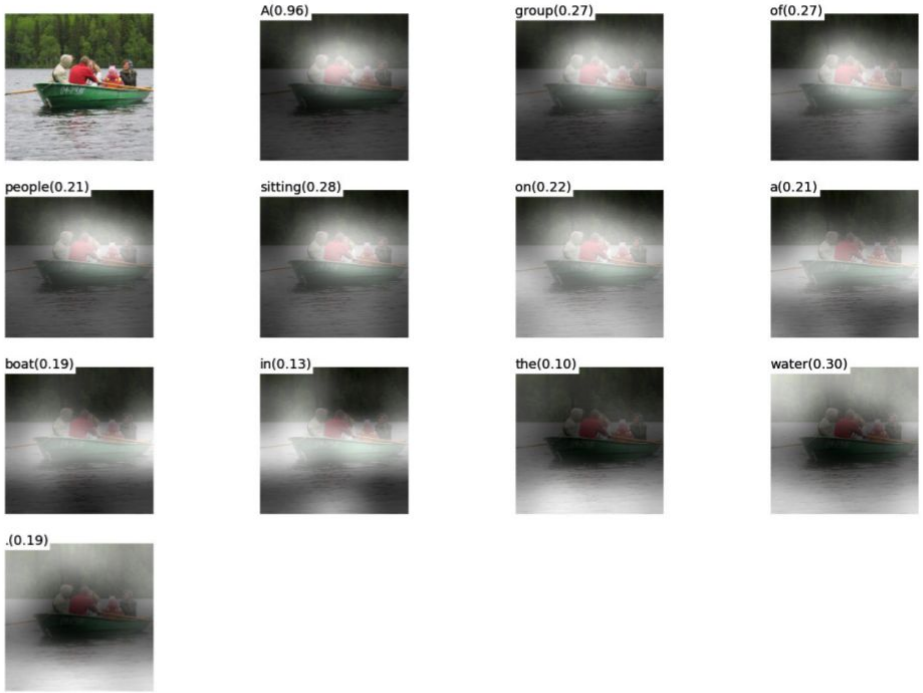
Recurrent Mixture Density Network for Spatiotemporal Visual Attention - L. Bazzani et al. (2016)

Implicit Attention Training



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention - K. Xu et al. (2015)

Implicit Attention Training



A group of people sitting on a boat in the water.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention - K. Xu et al. (2015)

Implicit Attention Training

A woman holding a clock in her hand.

METEOR metric

Flickr30k dataset

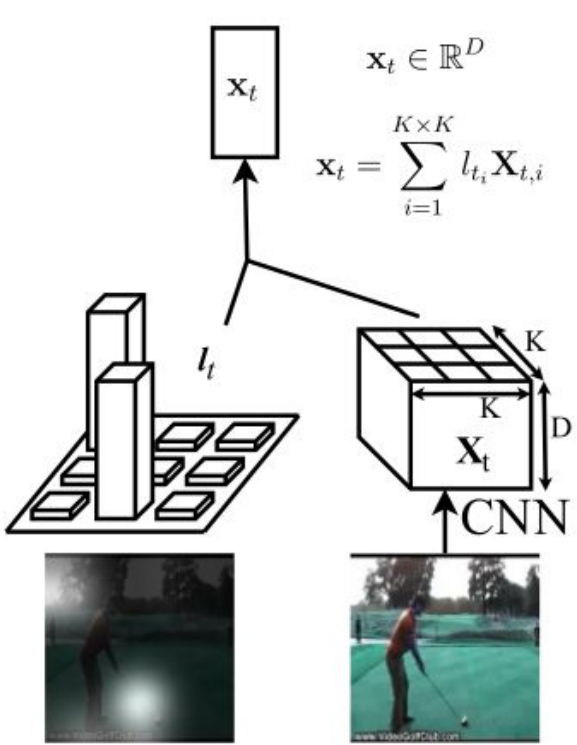
baseline: 16.88
soft-attention: 18.49 (+1.61)

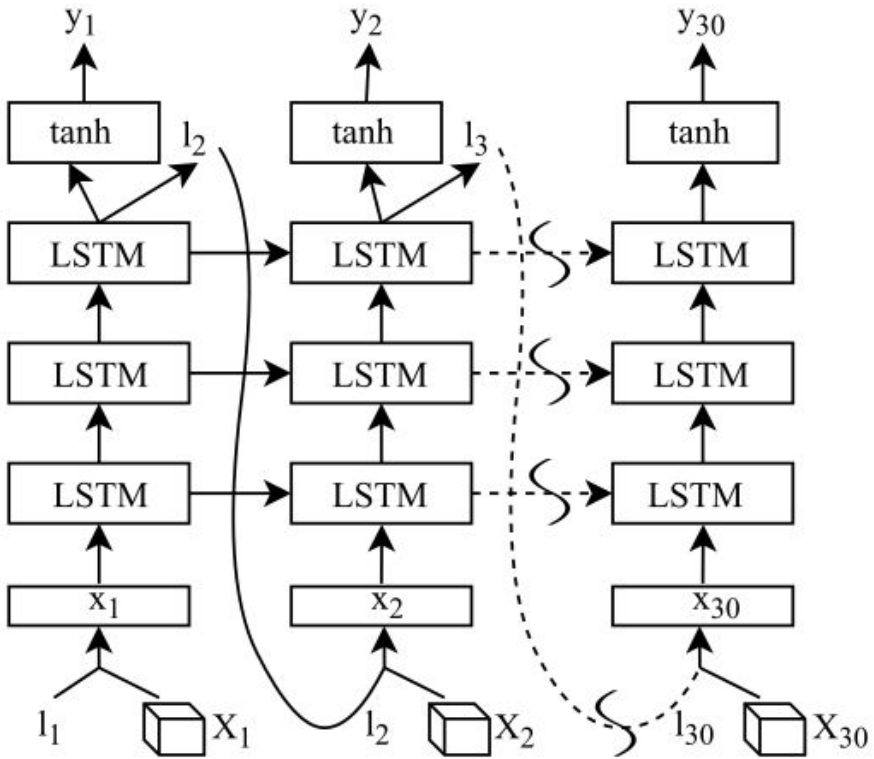
COCO dataset

baseline: 20.03
soft-attention: 23.9 (+3.87)

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention - K. Xu et al. (2015)

Implicit Attention Training

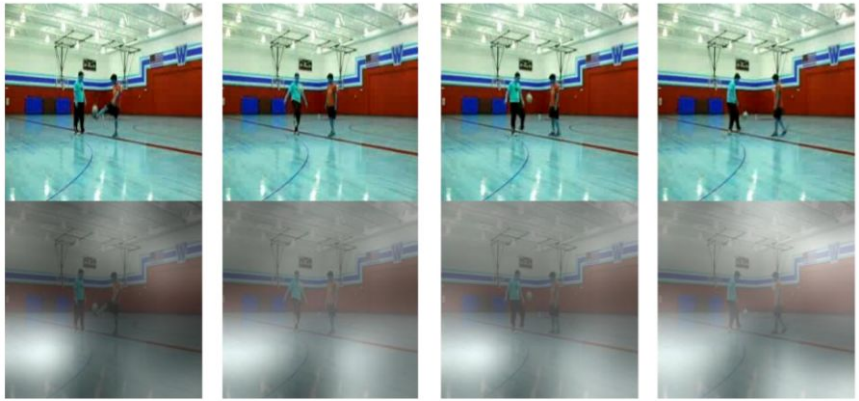




Action Recognition using Visual Attention - S. Sharma et al. (2015)

Implicit Attention Training

(a) Correctly classified as “cycling”



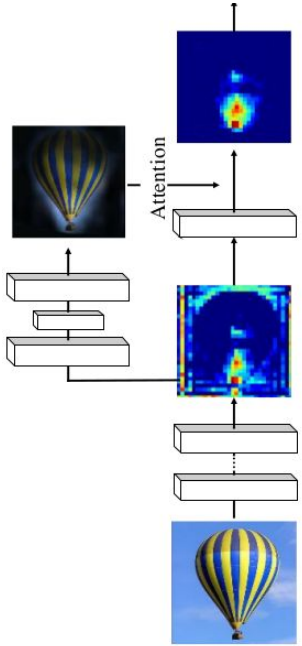
(a) Incorrectly classified as “diving”

UCF-11 dataset
baseline: 82.6%
attention: 85% (+2.4%)

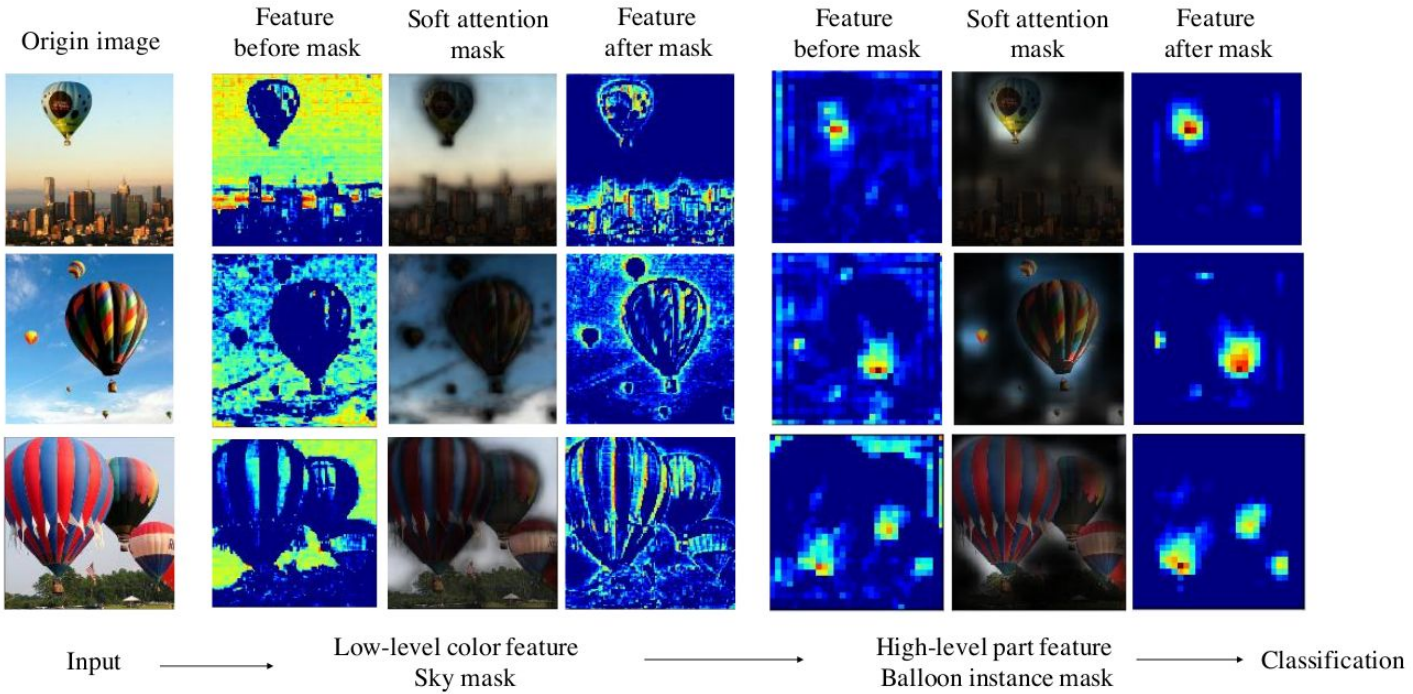
HMDB-51 dataset
baseline: 40.5%
attention: 41.3% (+0.8%)

Action Recognition using Visual Attention - S. Sharma et al. (2015)

My Research - 3D Residual Attention

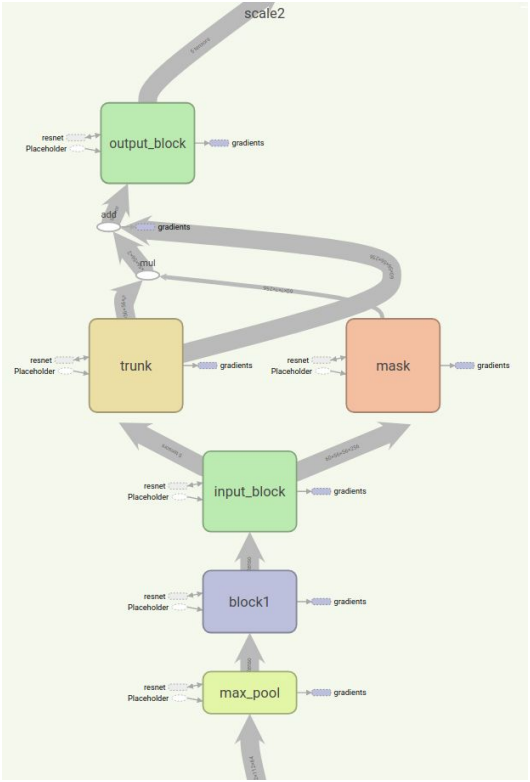


Attention mechanism

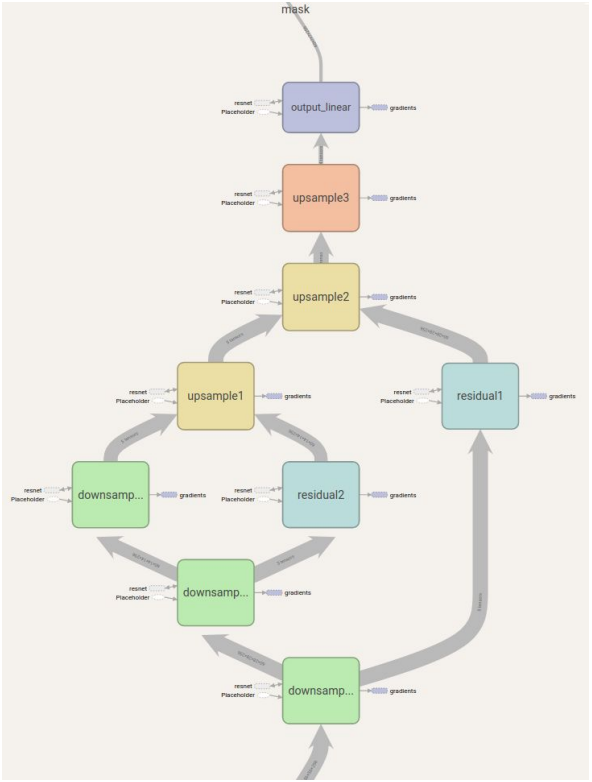


Residual Attention Network for Image Classification - F. Wang et al. (2017)

My Research - 3D Residual Attention

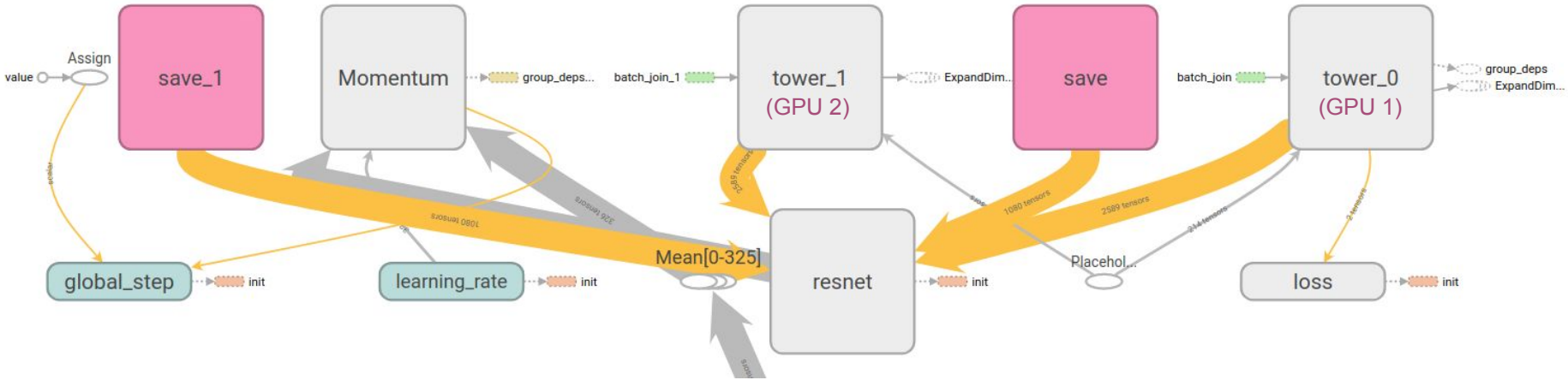


Attention Block



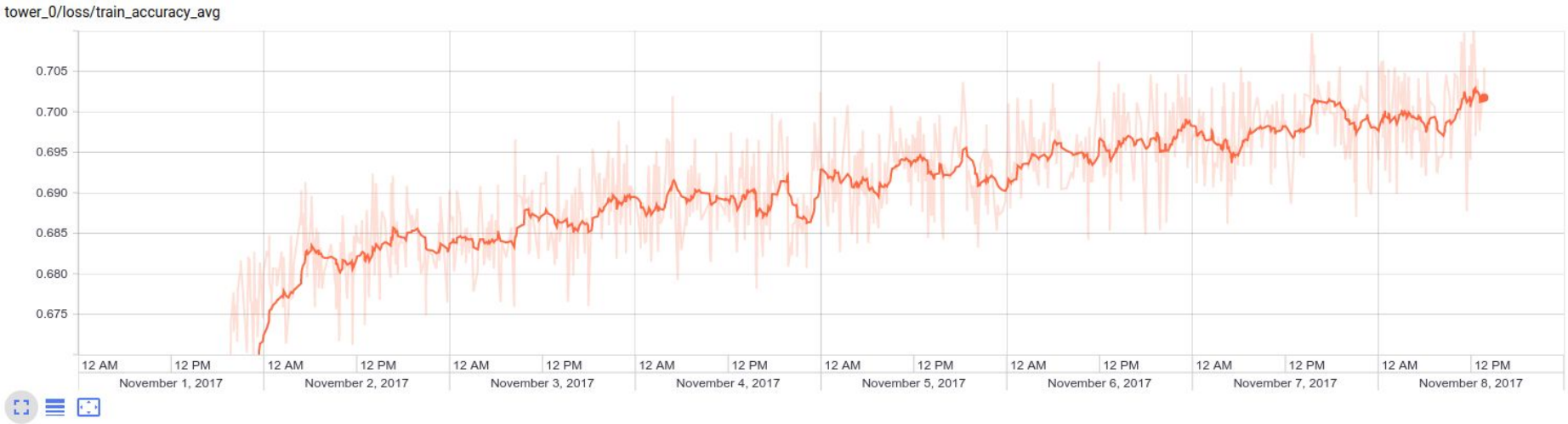
Attention Module

My Research - 3D Residual Attention



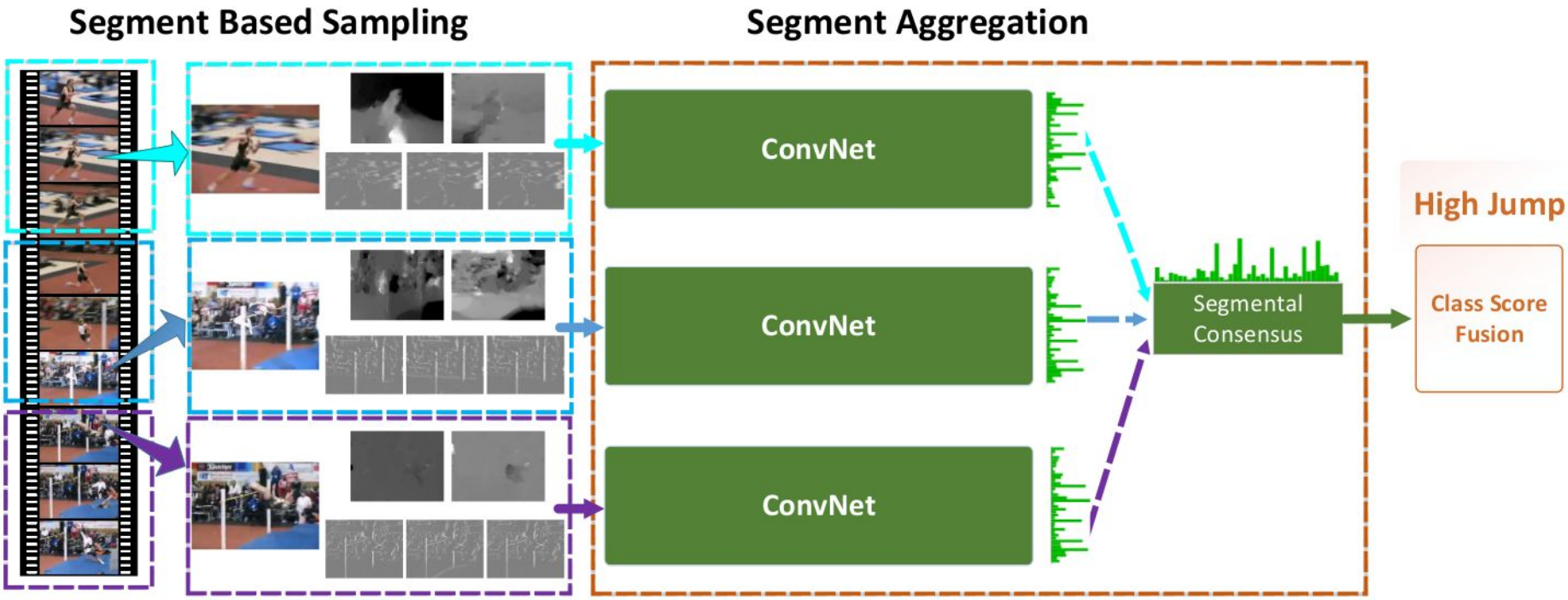
Large 3D ConvNets need to be trained across multiple GPUs ...

My Research - 3D Residual Attention



… and for a long time.

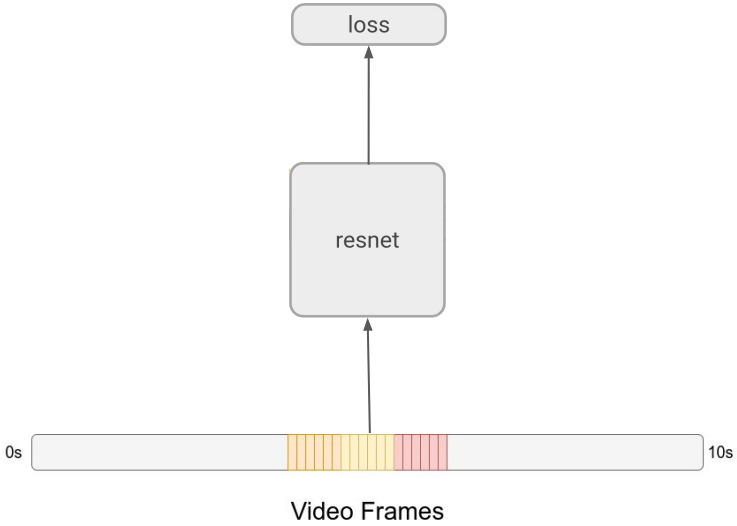
My Research - Temporal Segments



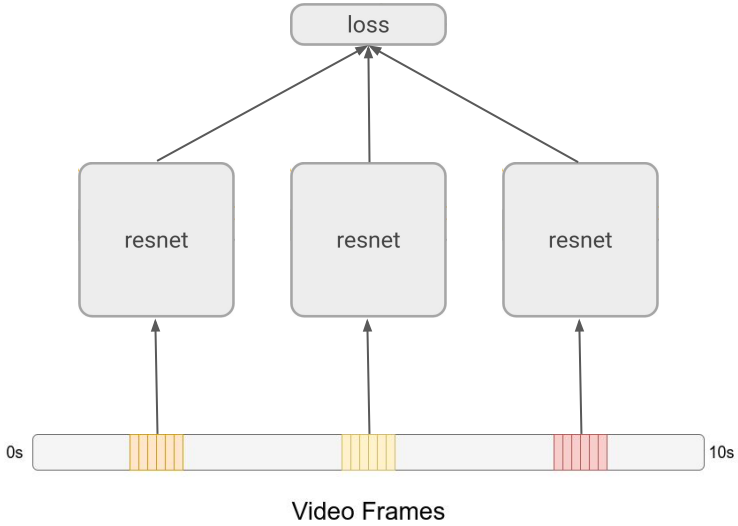
Temporal Segment Networks: Towards Good Practices for Deep Action Recognition - L. Wang et al. (2016)

My Research - Temporal Segments

Model single **large stack** of frames



Model several smaller **temporal segments**



Observations

- ConvNets with attention need **longer training** time but achieve **better performance**
- when presented with global context (temporal segments), ConvNets tend to **overfit** => strong regularization and a lot of training data are needed
- many actions are difficult to recognize from video but easily identified from **sound** - there is too much focus on understanding video

Conclusion

- Modelling videos is challenging for several reasons:
  - Deep Convolutional Networks can only process a couple of frames at a time due to memory restrictions during the **training** phase
  - Videos contain a lot of redundant information that confuse the models
  - Understanding movement is challenging due to camera motion
- Sophisticated **attention mechanisms** in the spatial and temporal domain address some of these issues
- We need a new class of neural networks or a new learning algorithms that are more efficient in order to model long-term dependencies in videos

References

In the order of appearance:

- Visualizing and Understanding Convolutional Networks - M.D.Zeiler et al. (2013)
- Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning - C. Szegedy et al. (2016)
- Deep Residual Learning for Image Recognition - K. He et al. (2015)
- Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset - J.Carreira et al. (2017)
- Beyond Short Snippets: Deep Networks for Video Classification - J.Y. Ng et al. (2015)
- Convolutional Two-Stream Network Fusion for Video Action Recognition - C. Feichtenhofer et al. (2016)
- Learning Spatiotemporal Features with 3D Convolutional Networks - D. Tran et al. (2014)
- Separate visual pathways for perception and action - M.A. Goodale et al. (1992)
- Recurrent Mixture Density Network for Spatiotemporal Visual Attention - L. Bazzani et al. (2016)
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention - K. Xu et al. (2015)
- Action Recognition using Visual Attention - S. Sharma et al. (2015)
- An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data - S. Song et al. (2016)
- Residual Attention Network for Image Classification - F. Wang et al. (2017)
- Temporal Segment Networks: Towards Good Practices for Deep Action Recognition - L. Wang et al. (2016)

ConvNet visualizations were created using Tensorboard, a tool included in the Tensorflow deep learning library.

Acknowledgements

I would like to thank Showmax for supporting my research and doc. Pavel Kordík for reviewing this presentation.