VideoLSTM

Convolves, attends and flows for action recognition

Zhenyang Li

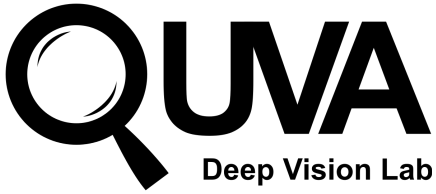
Kirill Gavrilyuk

Efstratios Gavves

Mihir Jain

Cees Snoek

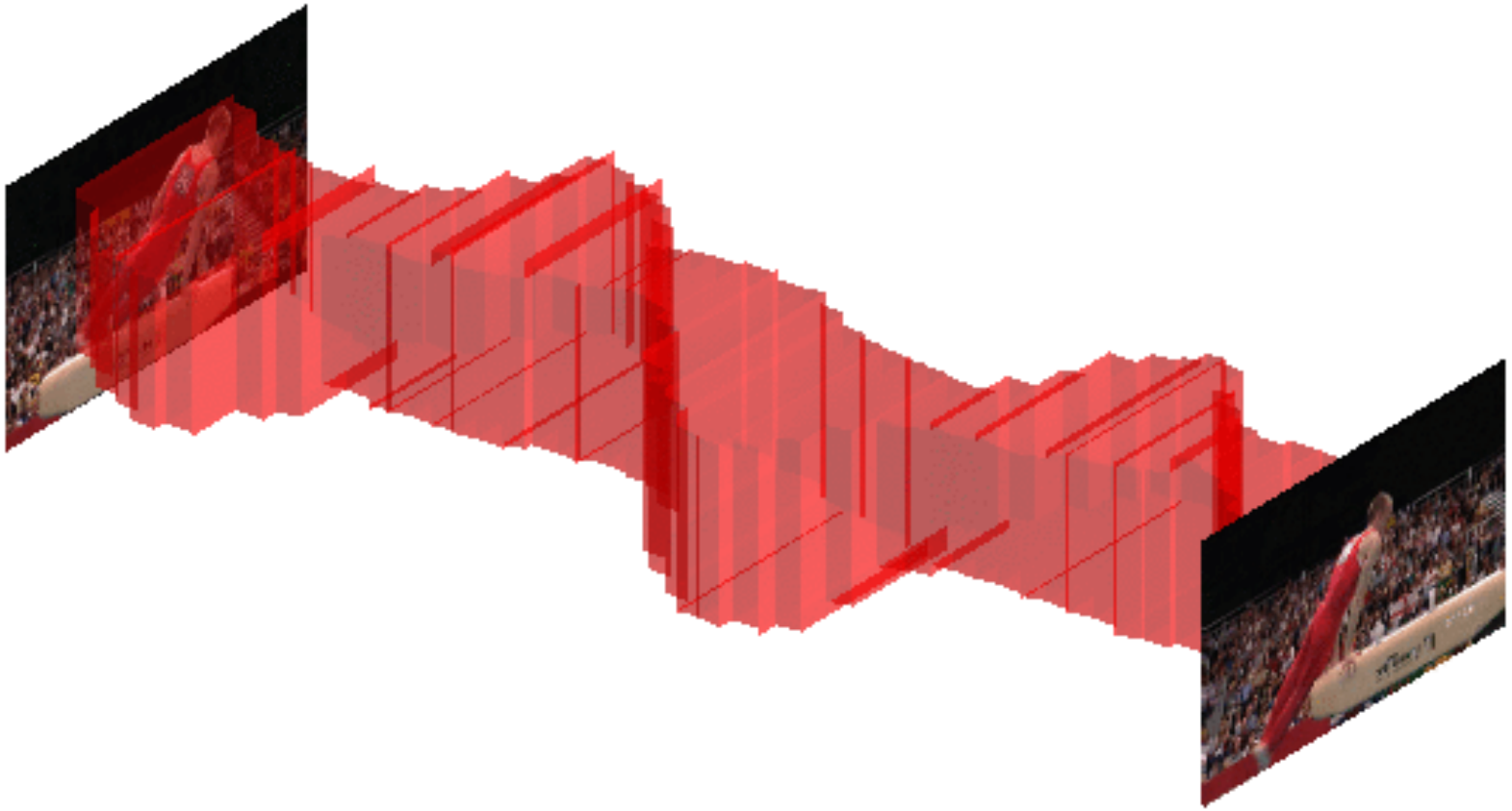
University of Amsterdam
The Netherlands



Motivation: Internet of things that video

Goal: Action Recognition

Understand **what** is happening **where** and **when**



Related work

DEEP LEARNING FOR ACTIONS

ConvNet

Ji et al. ICML10

Karpathy et al., CVPR14

Simonyan & Zisserman, NIPS14

Tran et al., ICCV15

3D convolutions

Need large amounts of data to learn filters

Two-stream

Learn spatial and temporal filters separately

We propose a more principled approach for learning frame-to-frame appearance and motion transitions.

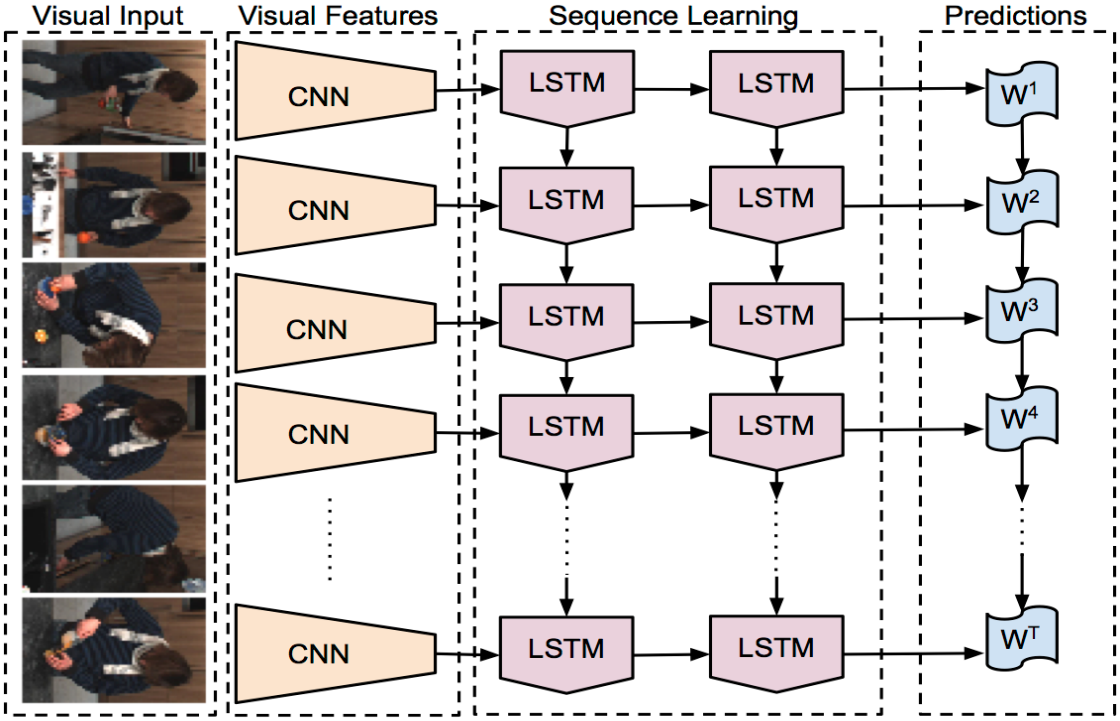
We localize the action as well.

Baccouche et al. ICANN10
Donahue et al. / Ng et al. CVPR15
Srivastava et al. ICML15

LSTM

LSTM models sequential memories in the long and short term

Use ConvNet fc **vectors** as input, no spatial information encoded



Rather than vectorizing a video frame, we rely on **convolutions**

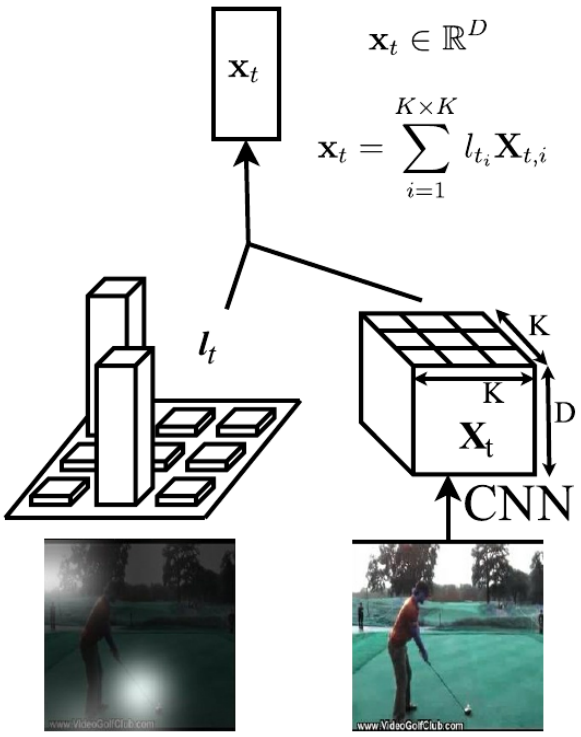
Sharma et al. NIPS15 / ICLR16

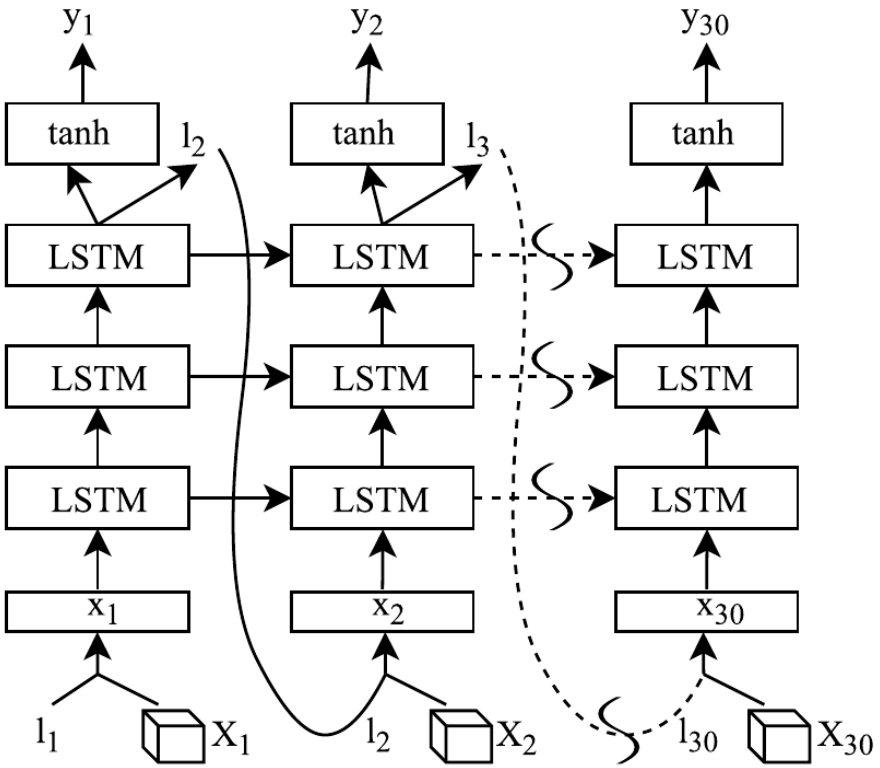
A(ttention)LSTM

Look for best locations leading to correct action classification

Stays close to soft-Attention architecture for image captioning [Xu et al. ICML15],

Vectorizes attention and appearance, and ignores the motion inside a video.





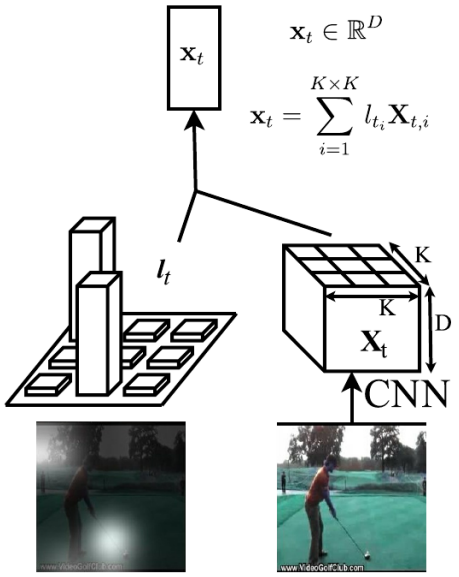
Sharma et al. NIPS15 / ICLR16

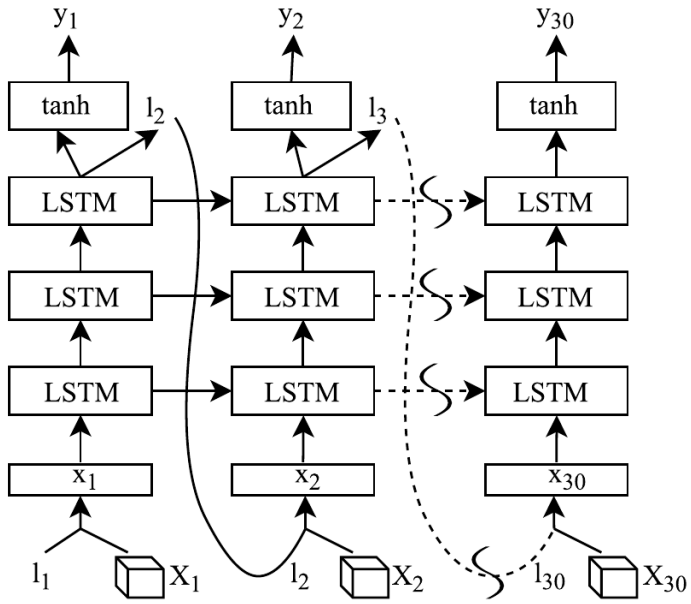
A(ttention)LSTM

Look for best locations leading to correct action classification

Stays close to soft-Attention architecture for image captioning [Xu et al. ICML15],

Vectorizes attention and appearance, and ignores the motion inside a video.





We add **convolutions** and **motion** for better action classification

We localize the action as well.

Our proposal: VideoLSTM

Model spatiotemporal dynamics of videos by

- preserving spatial structure of the frames over time
- adding motion-based attention
- enabling action localization from action class labels only

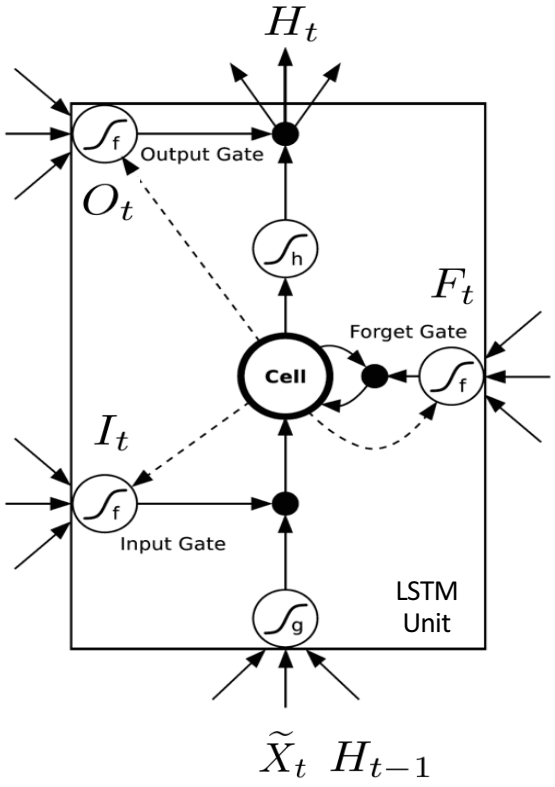
VIDEOLSTM

VideoLSTM convolves, attends and flows for action recognition.
Zhenyang Li, Efstratios Gavves, Mihir Jain, and Cees Snoek. Arxive16.
http://arxiv.org/abs/1607.01794

Convolutional (A)LSTM

Replace the fully connected multiplicative operations in an LSTM unit with convolutional operations

$$
\begin{aligned}
I_t &= \sigma(W_{xi} * \widetilde{X}_t + W_{hi} * H_{t-1} + b_i) \\
F_t &= \sigma(W_{xf} * \widetilde{X}_t + W_{hf} * H_{t-1} + b_f) \\
O_t &= \sigma(W_{xo} * \widetilde{X}_t + W_{ho} * H_{t-1} + b_o) \\
G_t &= \tanh(W_{xc} * \widetilde{X}_t + W_{hc} * H_{t-1} + b_c) \\
C_t &= F_t \odot C_{t-1} + I_t \odot G_t \\
H_t &= O_t \odot \tanh(C_t),
\end{aligned}
$$



Generate attention by shallow ConvNet instead of MLP

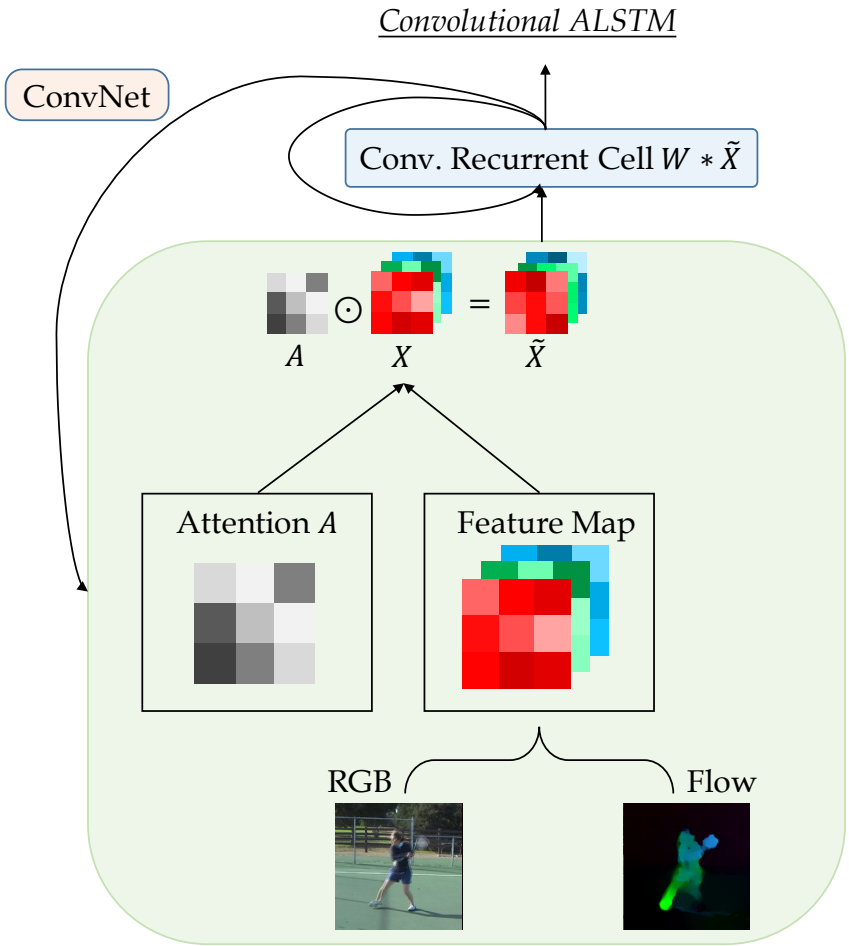
Convolutional ALSTM

Attention map is generated by a two-layer ConvNet

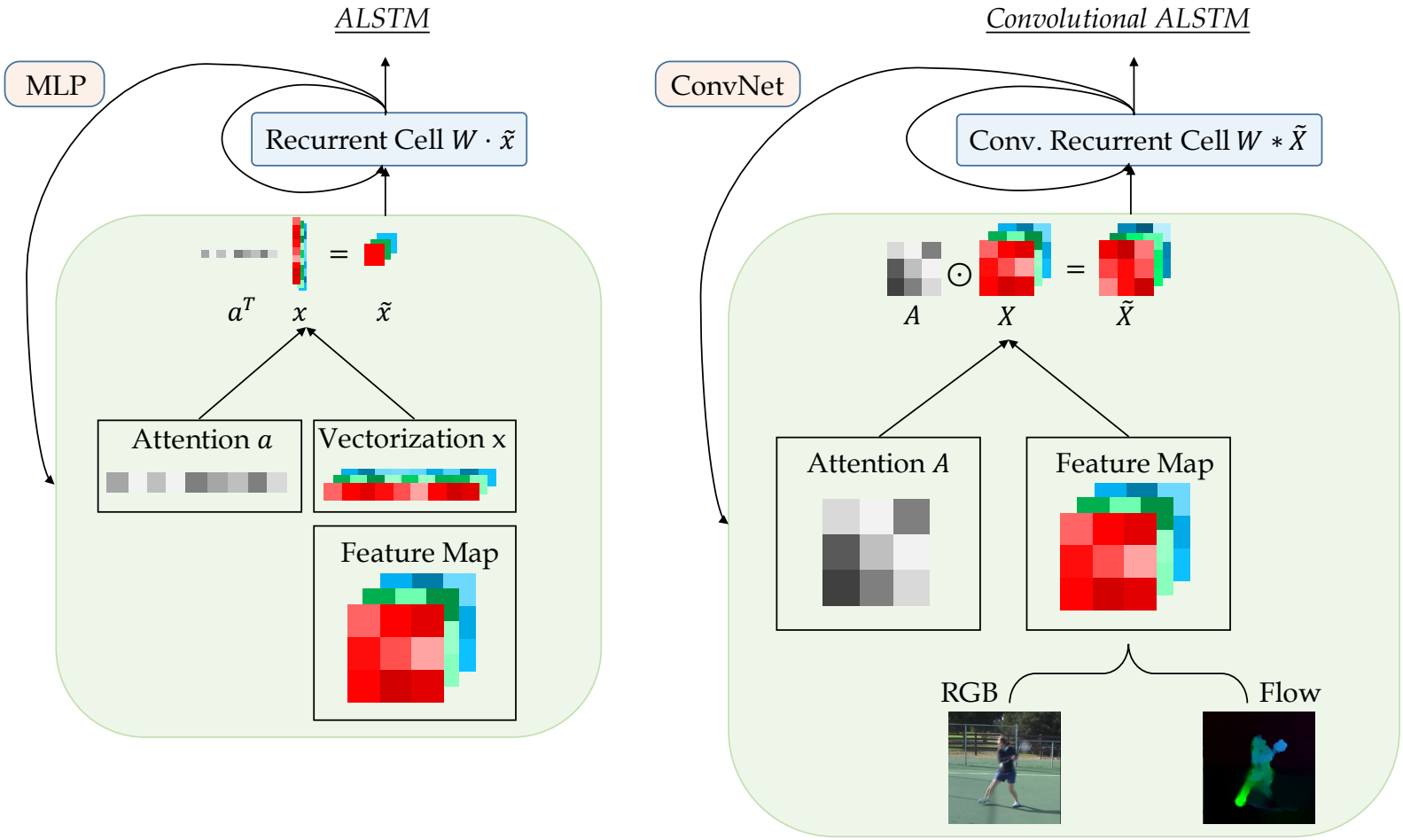
$$Z_t = W_z * \tanh(W_{xa} * X_t + W_{ha} * H_{t-1} + b_a)$$

$$A_t^{ij} = p(att_{ij}|X_t, H_{t-1}) = \frac{\exp(Z_t^{ij})}{\sum_i \sum_j \exp(Z_t^{ij})}$$

$$\widetilde{X}_t = A_t \odot X_t$$



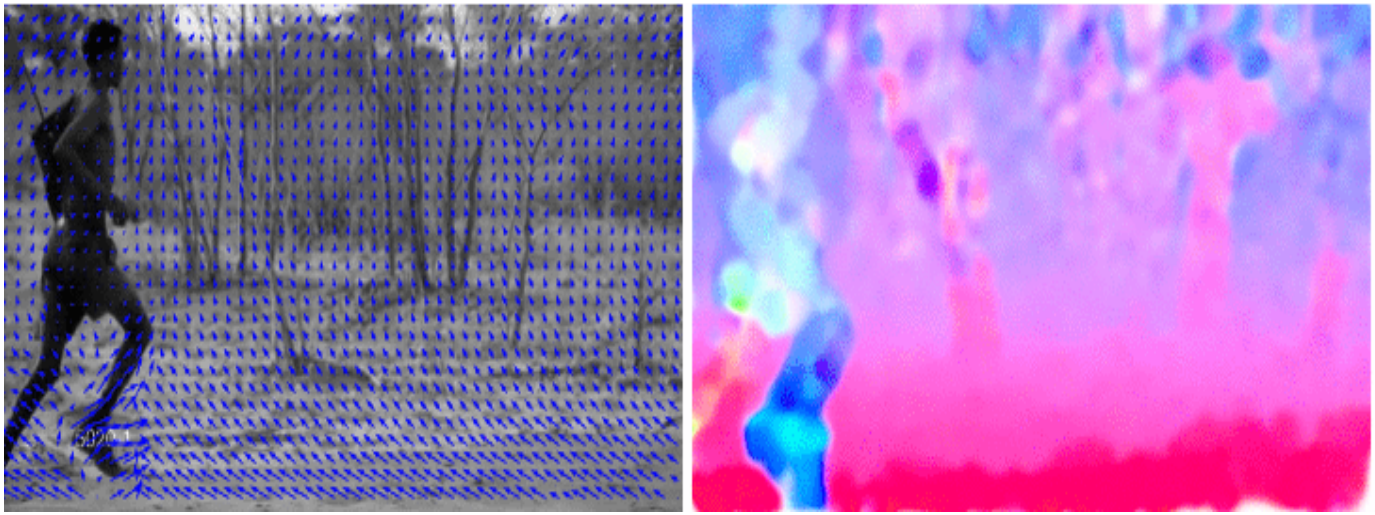
ALSTM vs Convolutional ALSTM



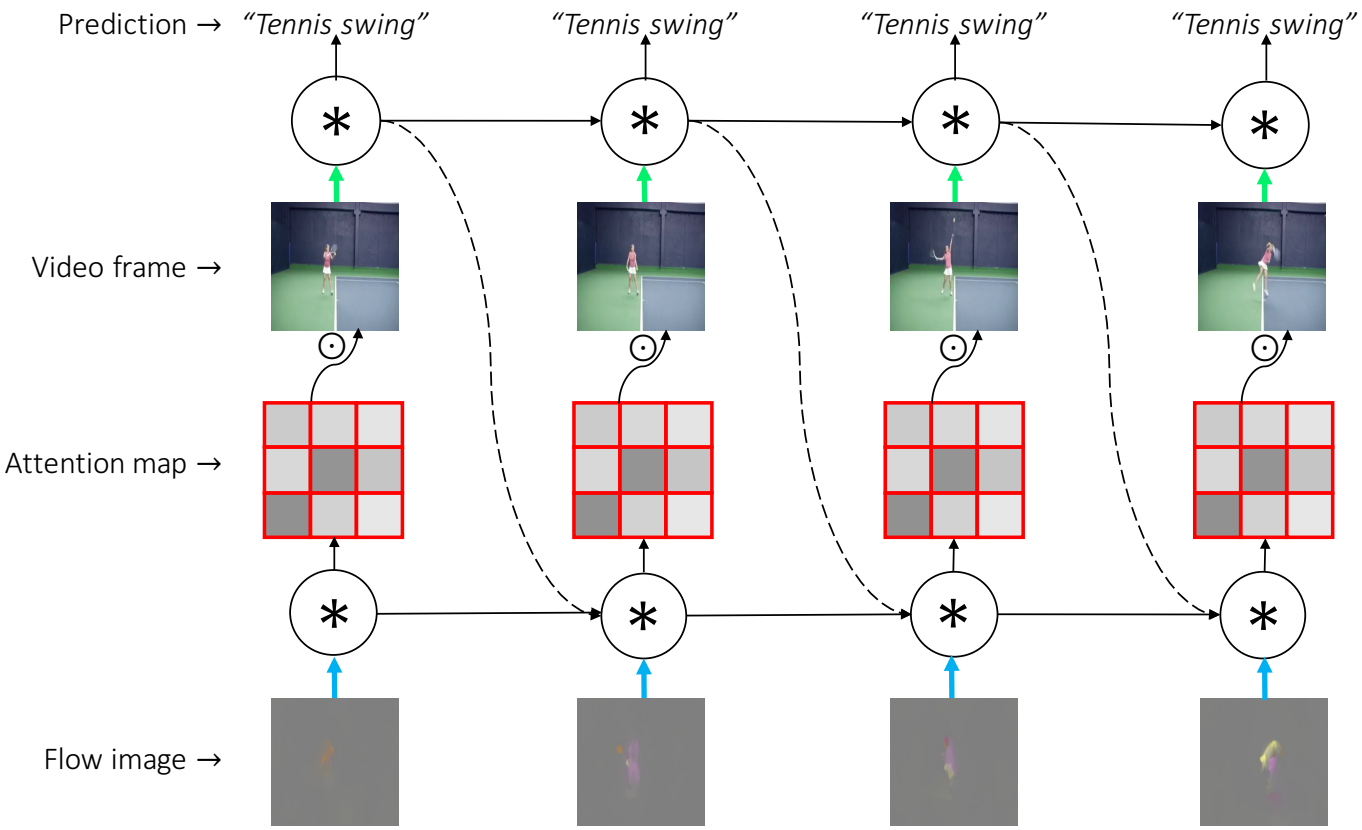
Convolutional ALSTM preserves spatial dimensions over time

Motion-based attention

Motion offers crucial clue where to attend in video



Motion-based attention



Motion information to infer the attention in each frame

Experimental setup

Datasets:

UCF101, **HMDB51** for action classification

Comparison using similar designs and training regime:

ConvNet: VGG-16 trained for both RGB frame and optical flow.

LSTM: Use subsequences of every 30 frames, extract fc7 or pool5 features at each frame as input.

Convolutions: 3x3 kernels for input-to-state and state-to-state transitions in LSTM, and 1x1 kernels to generate the attention map

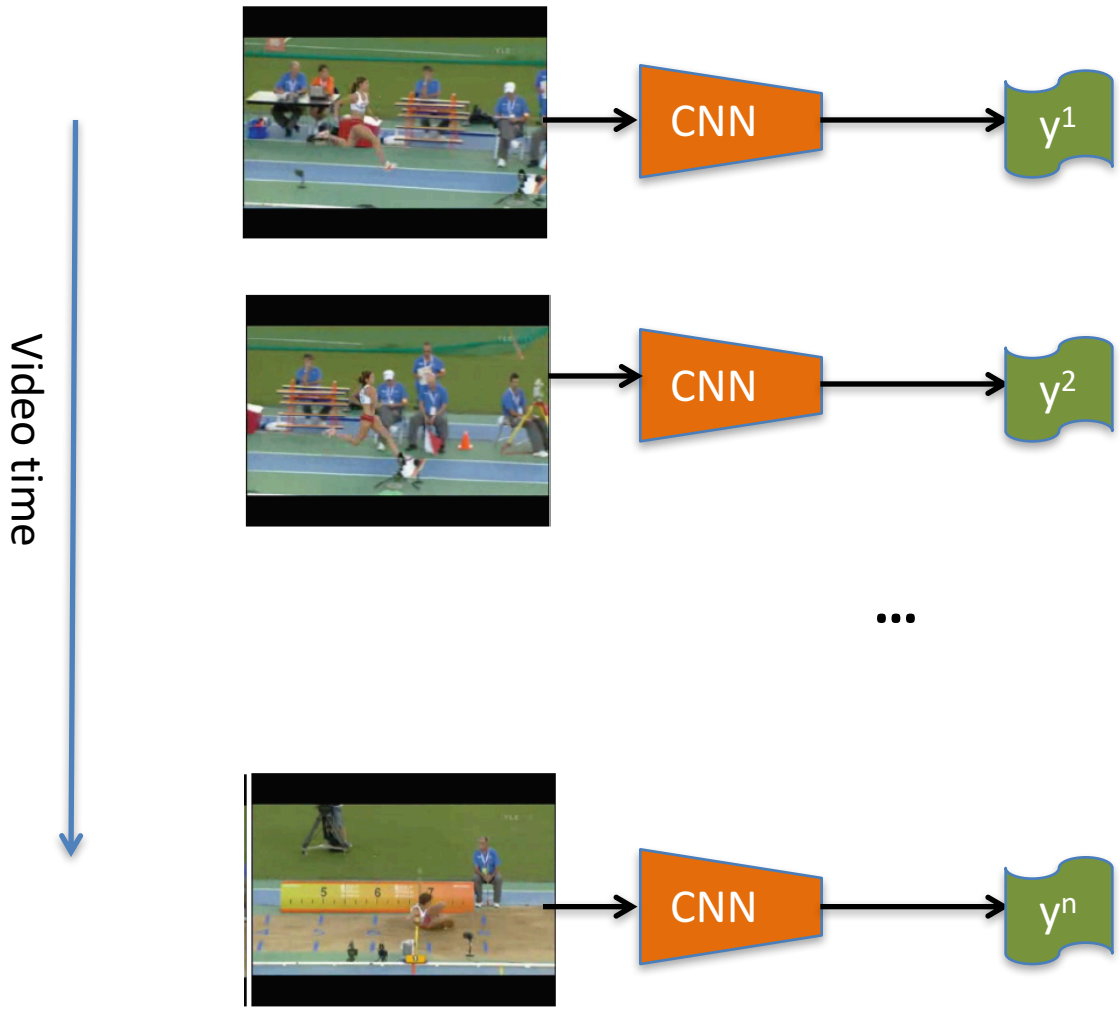
Experiments

1. What deep learning architecture?

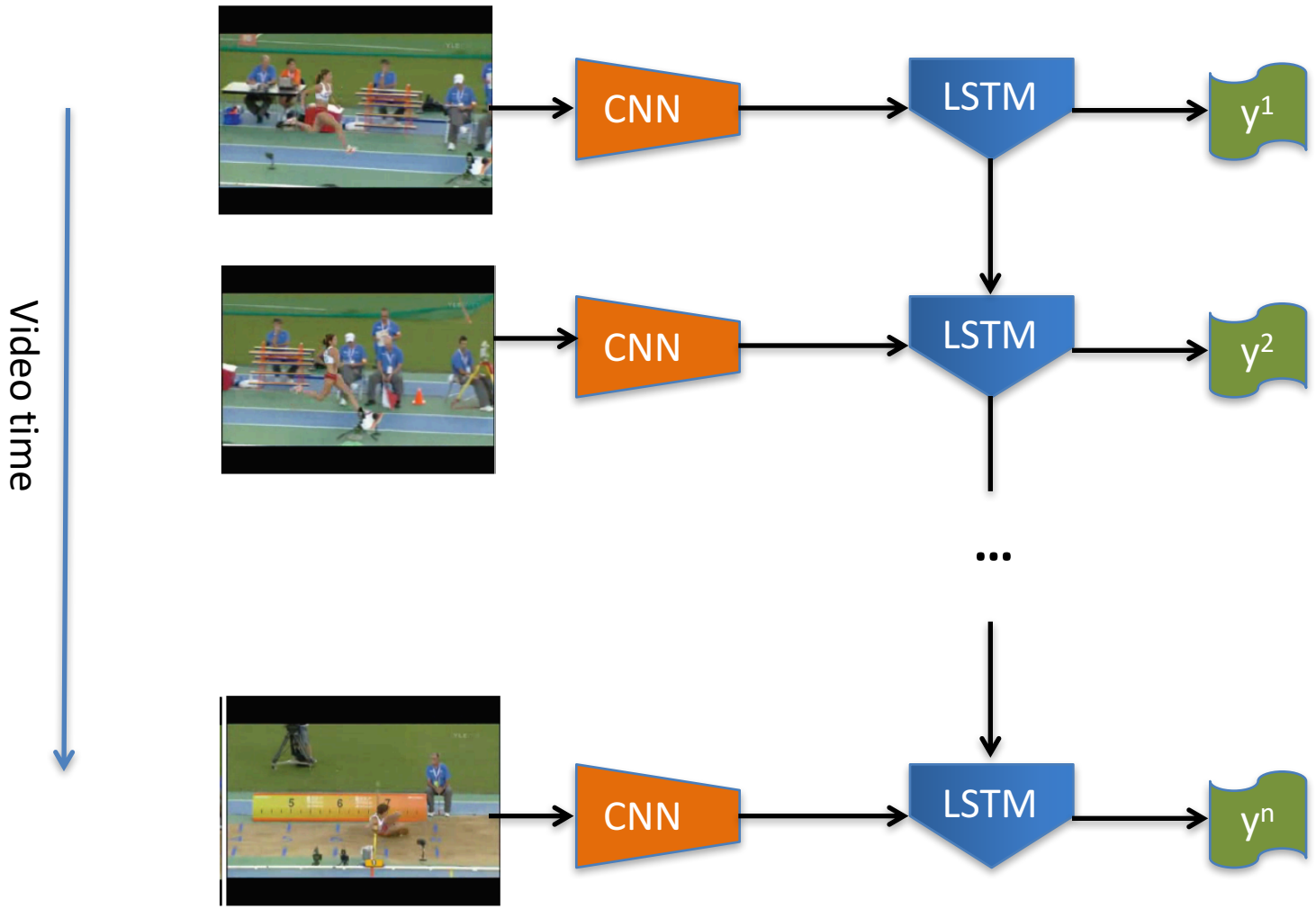
2. Influence of motion-based attention

3. Quality of action localization

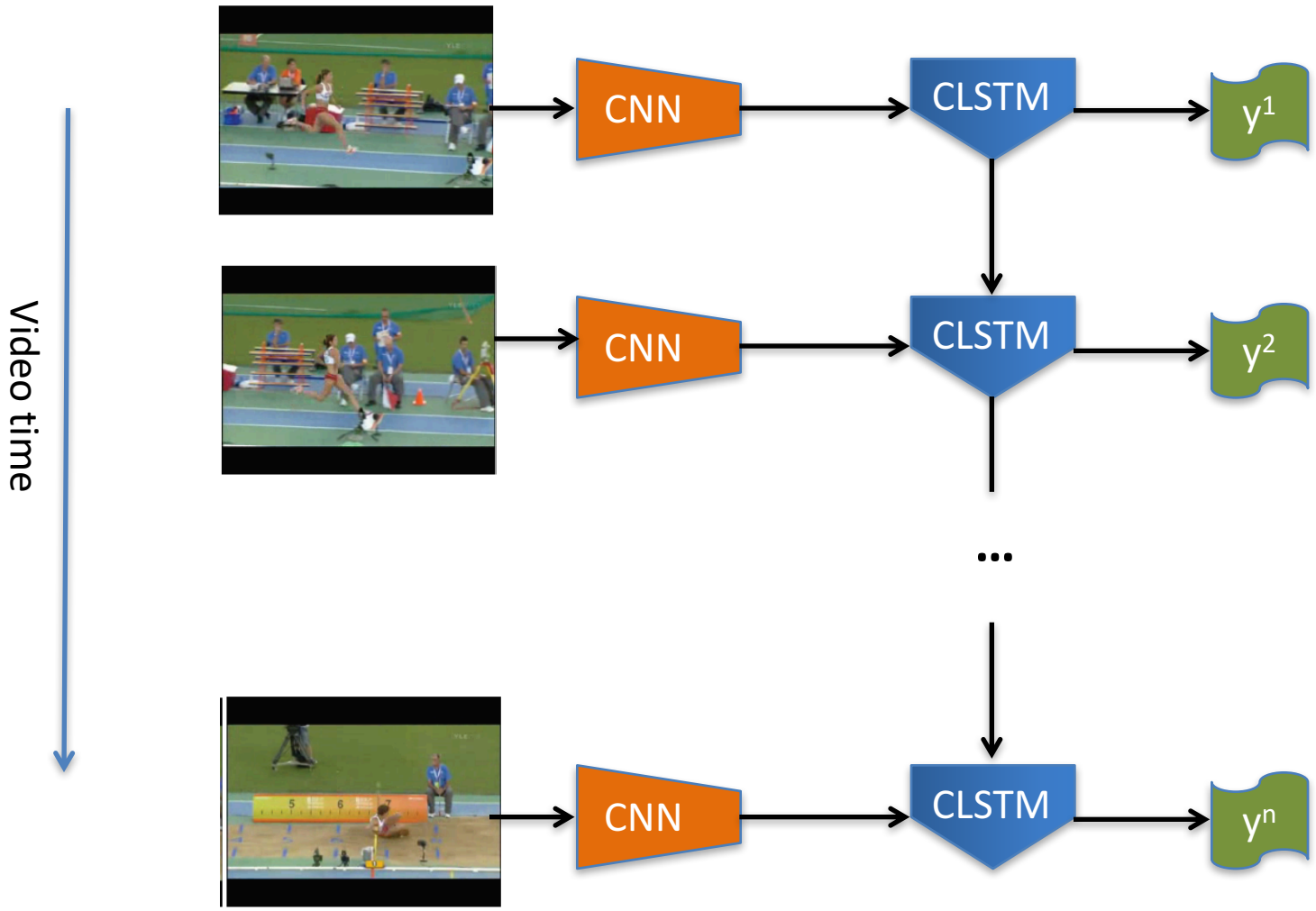
ConvNet



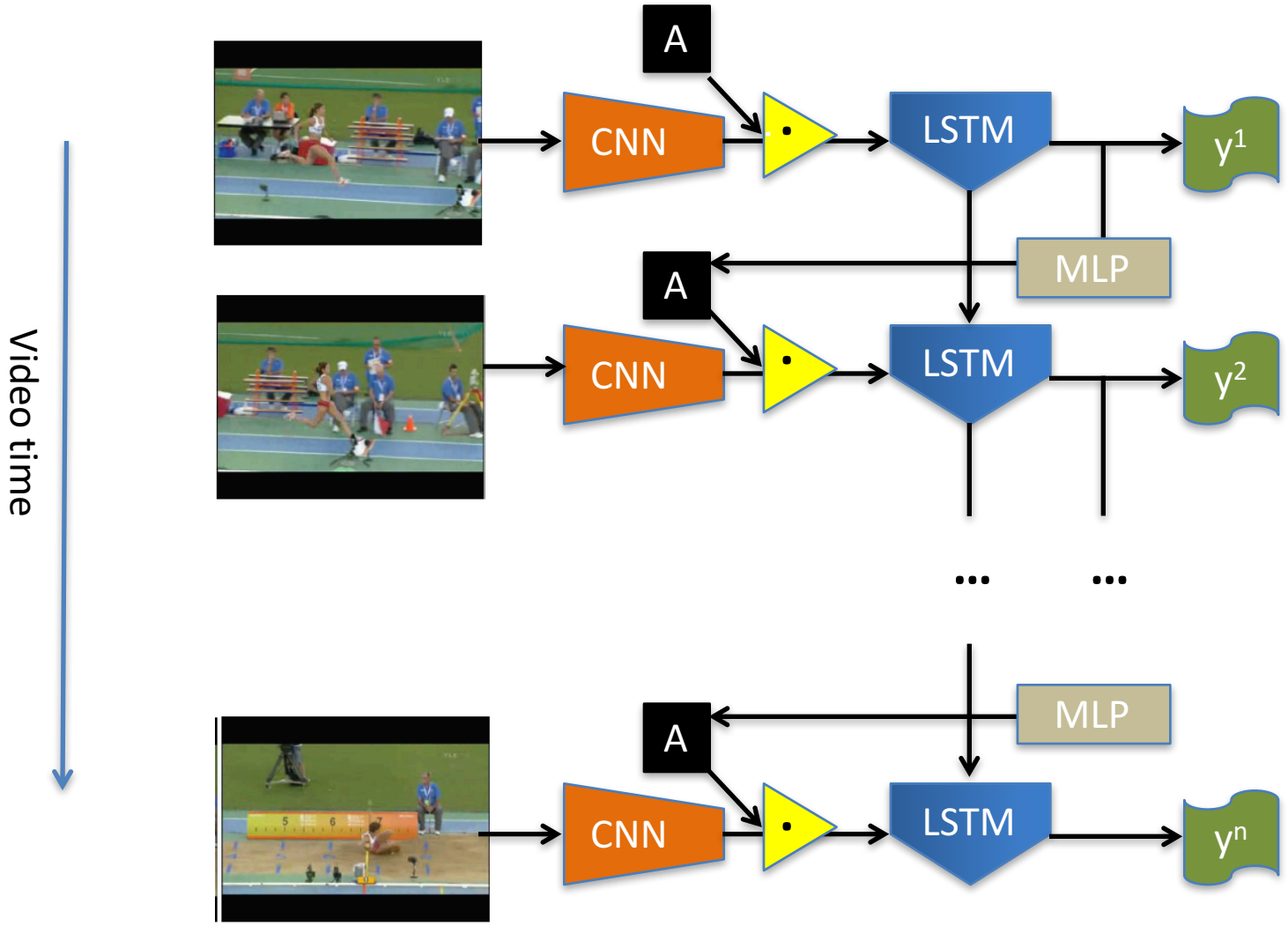
LSTM



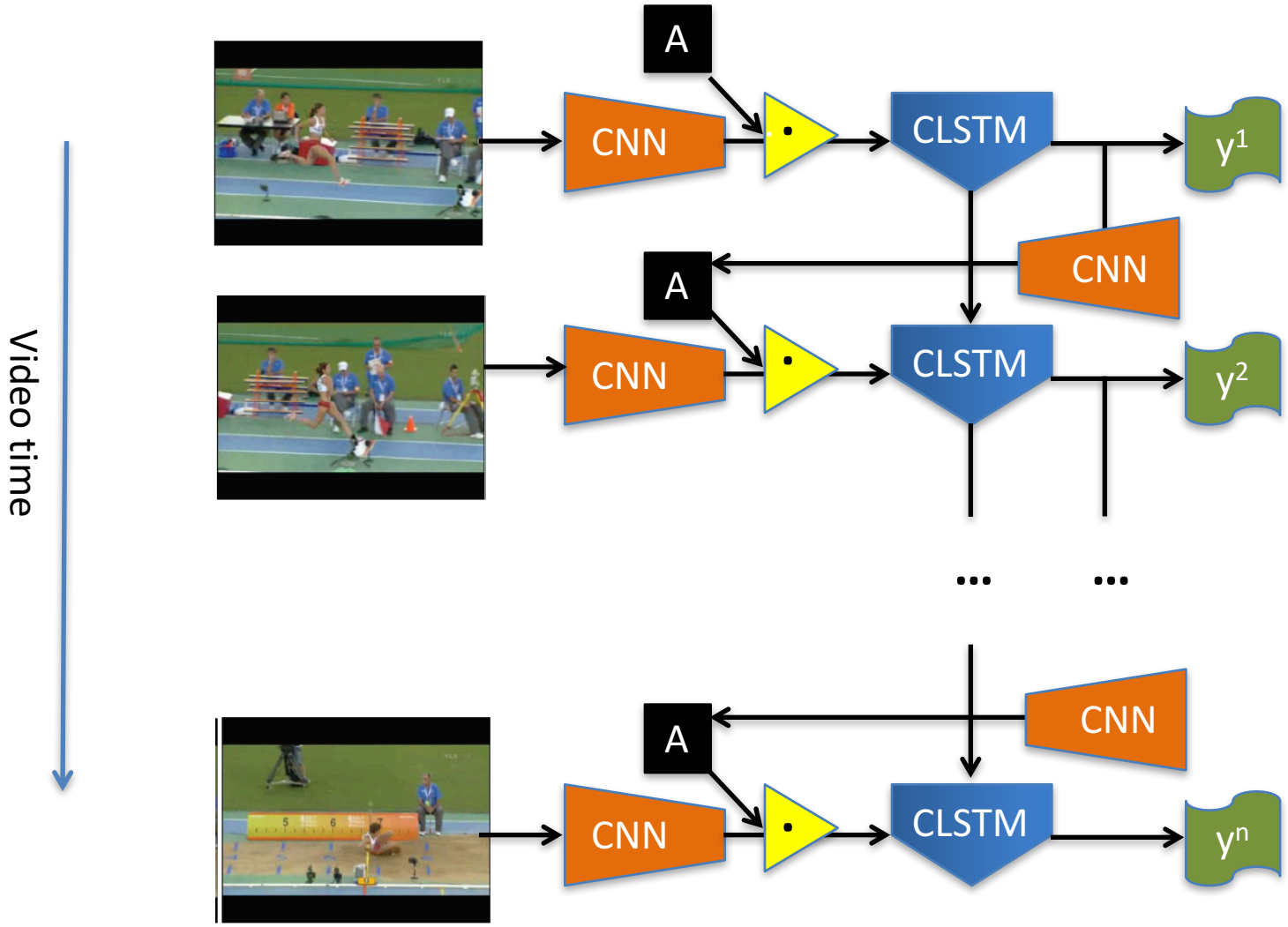
Convolutional LSTM



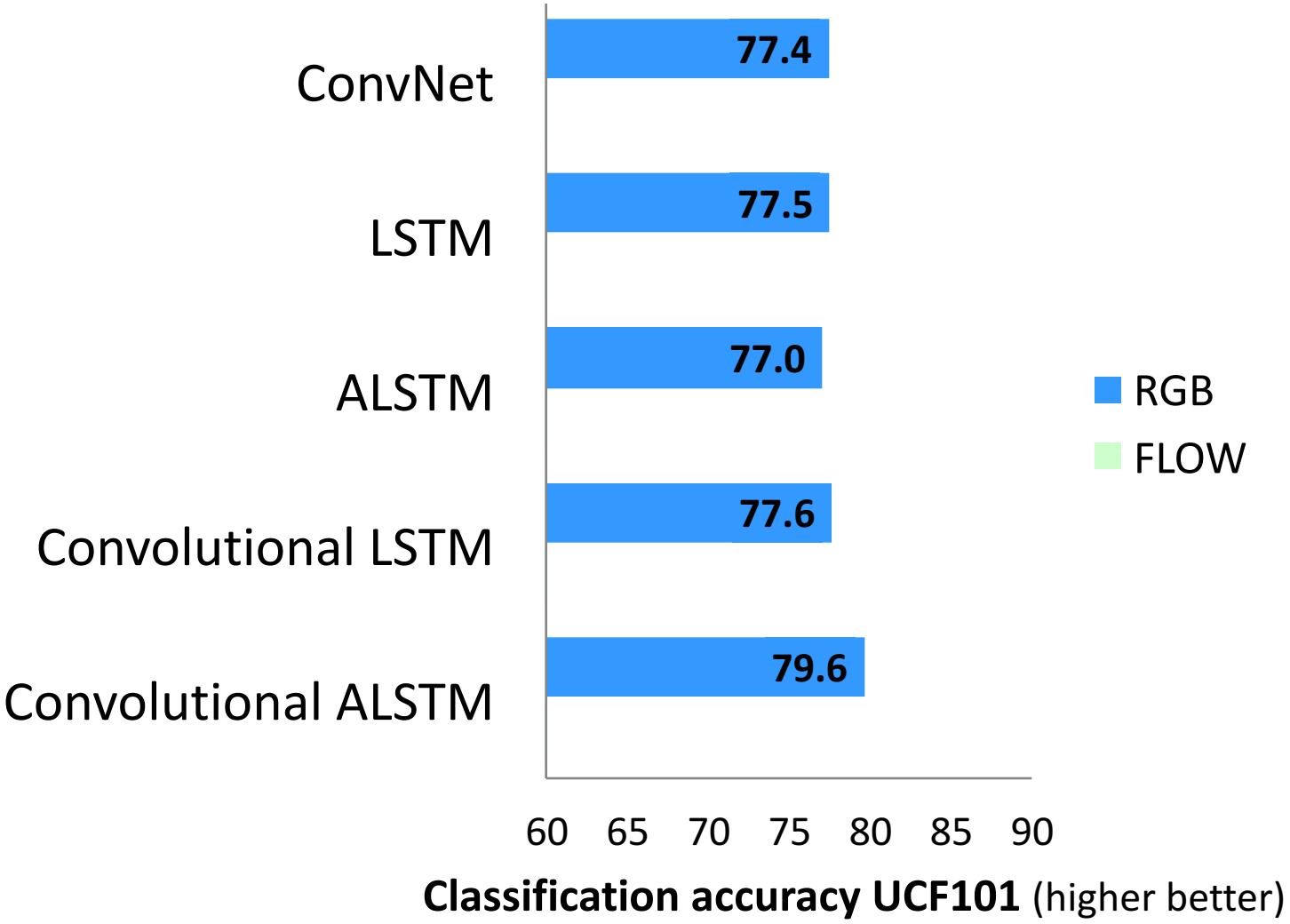
ALSTM



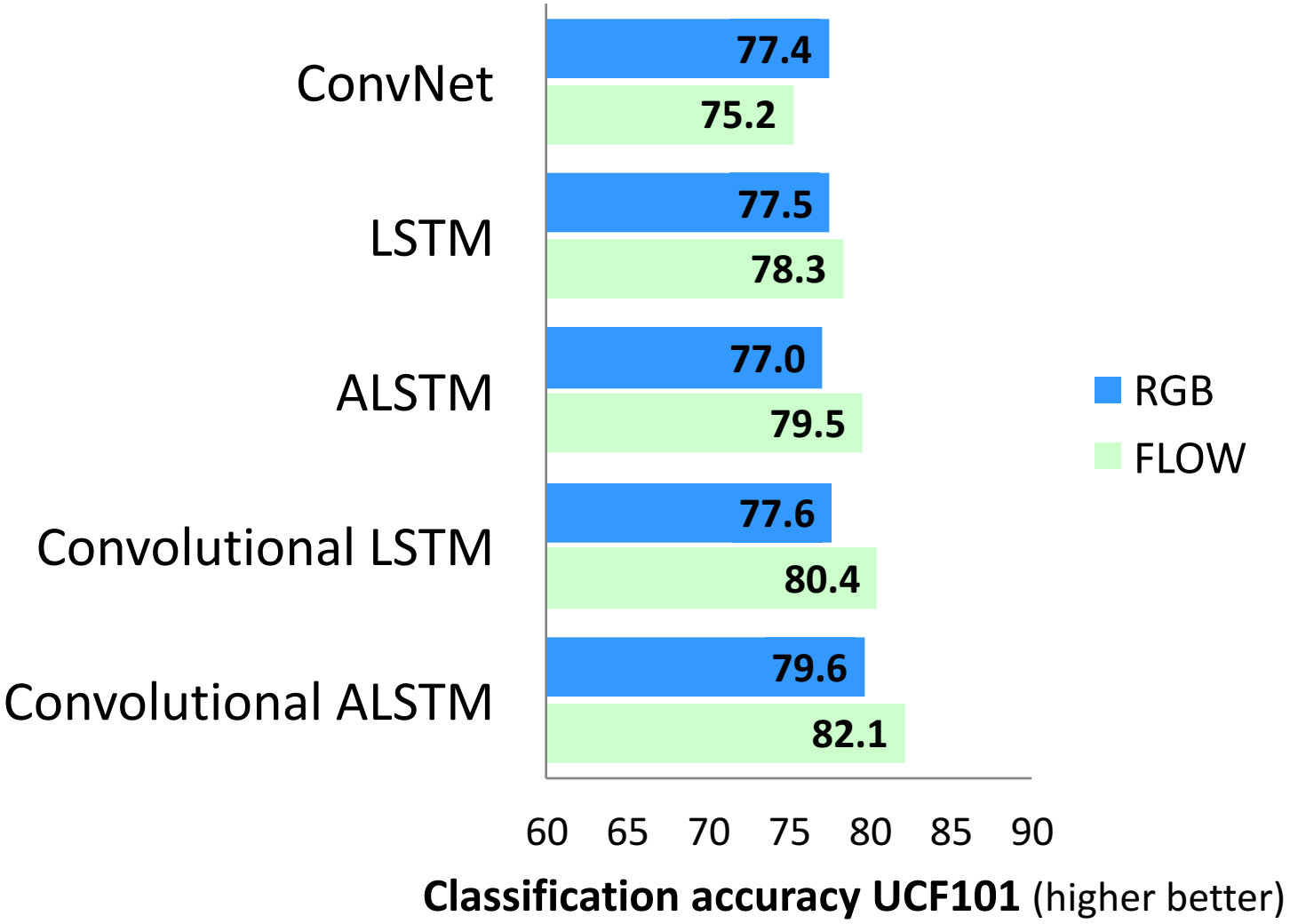
Convolutional ALSTM



Convolution, attention and flow



Convolution, attention and flow



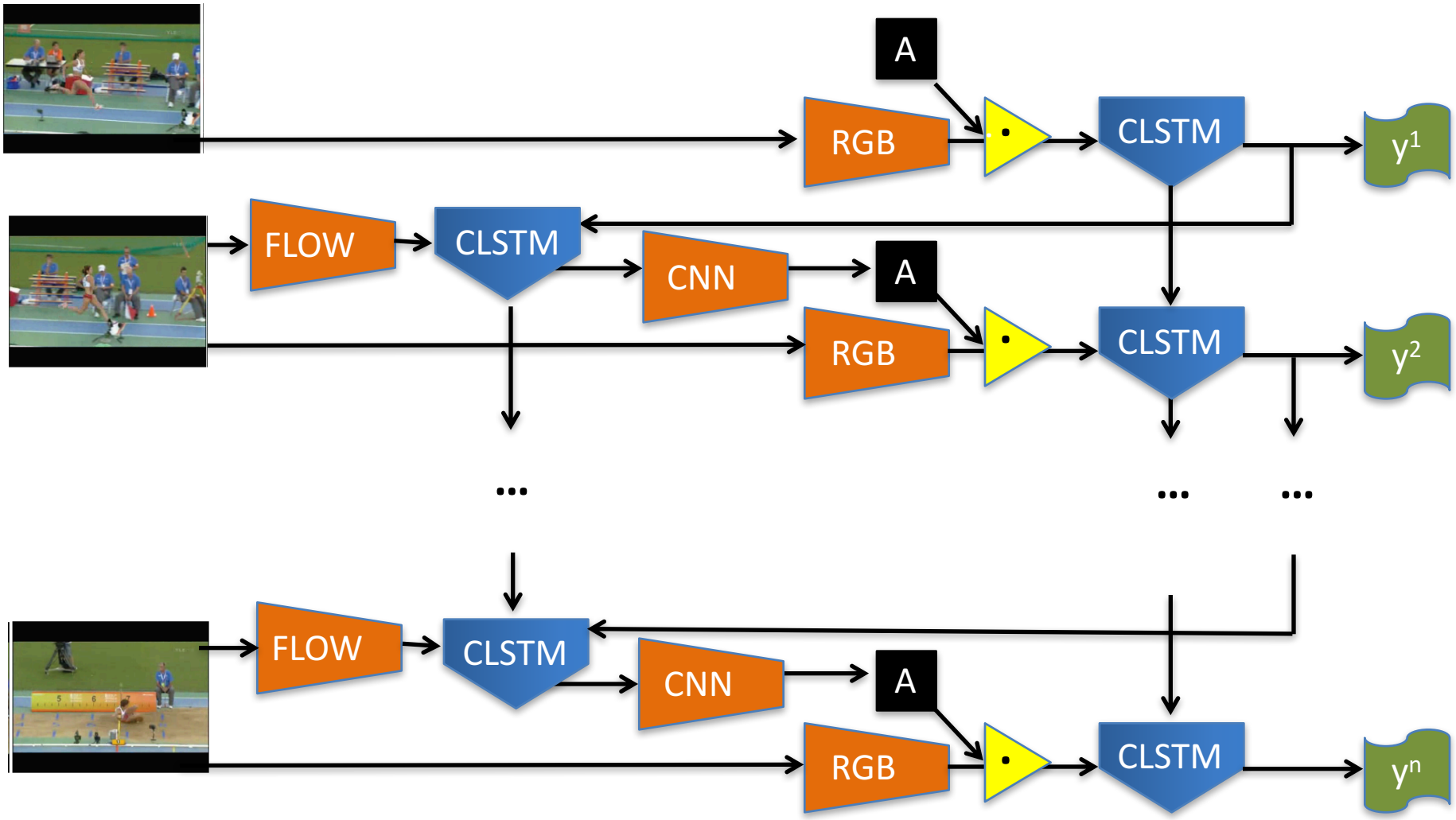
Experiments

1. What deep learning architecture?

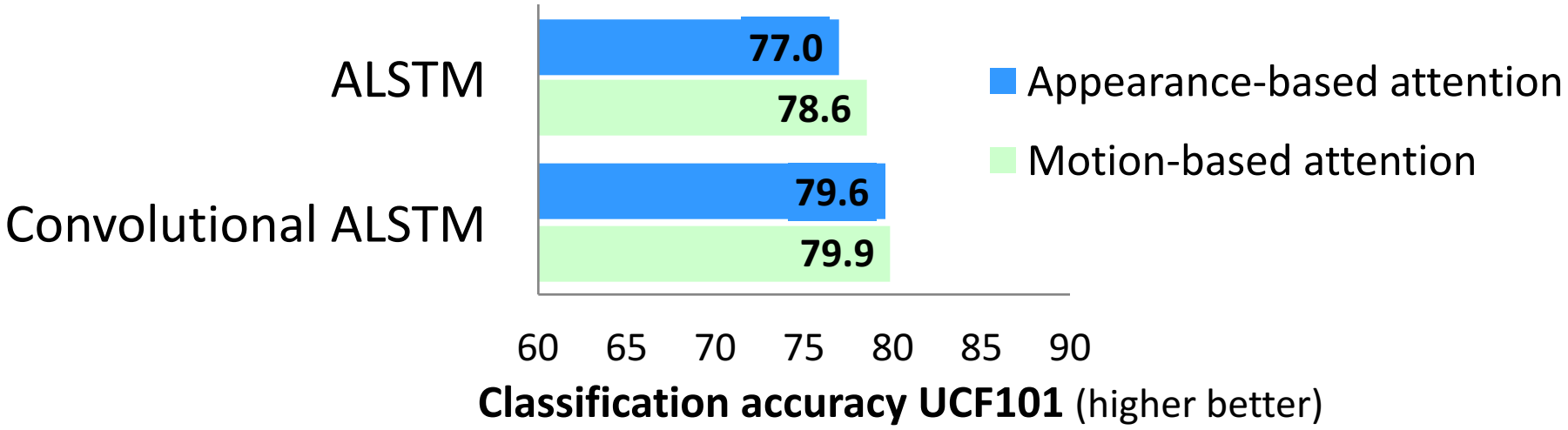
2. Influence of motion-based attention

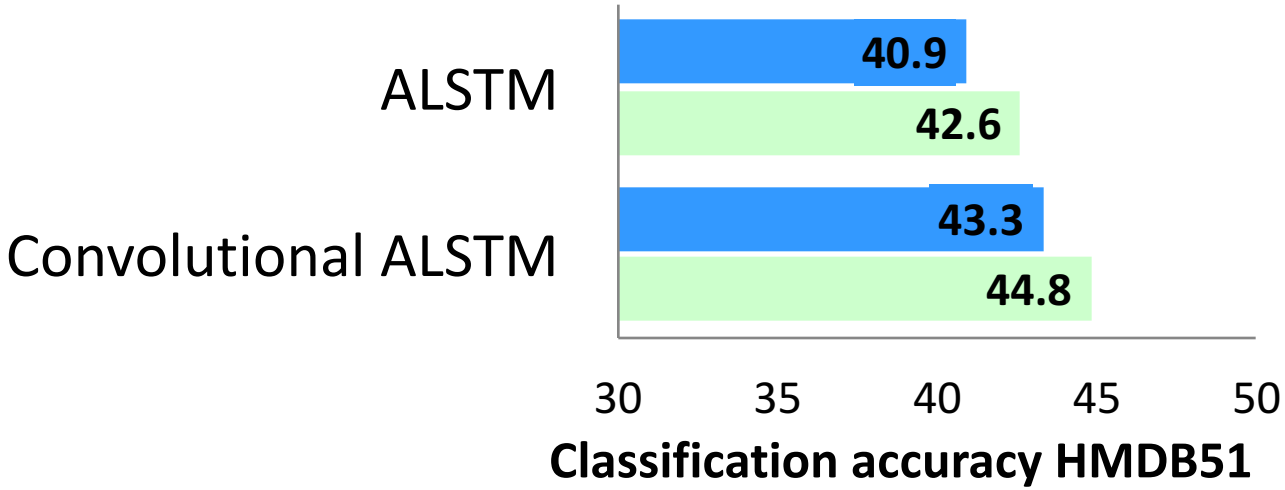
3. Quality of action localization

Recap: Motion-based attention



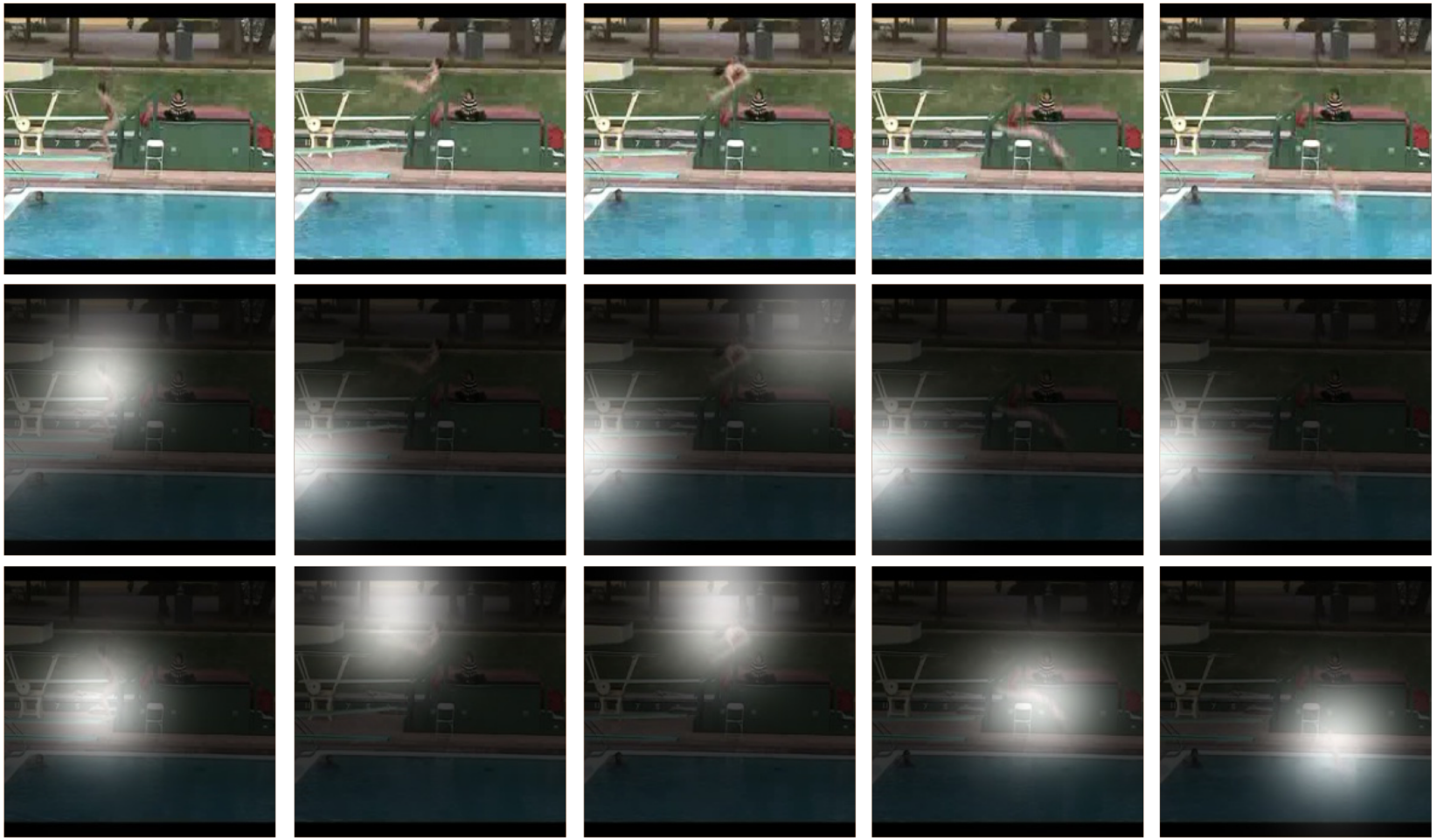
Motion attention makes more sense





Motion attention makes more sense

Motion attention makes more sense



Experiments

1. What deep learning architecture?

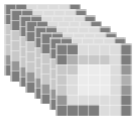
2. Influence of motion-based attention

3. Quality of action localization

Temporal smoothing

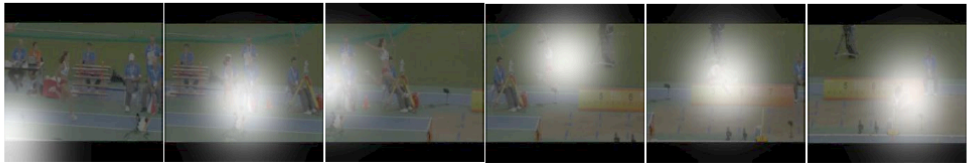
Input frame sequence

VideoLSTM



Attention feature maps

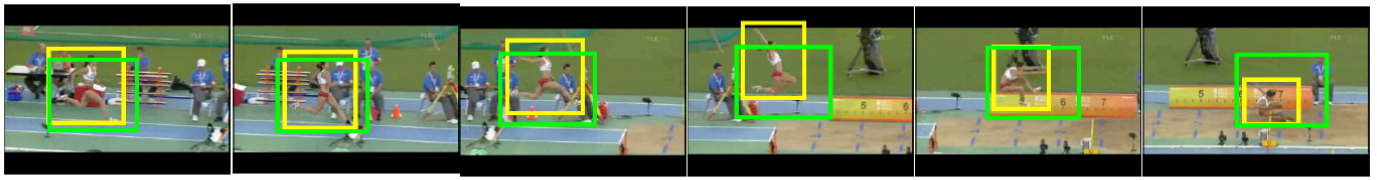
Up-sampling



Attention saliency over frames

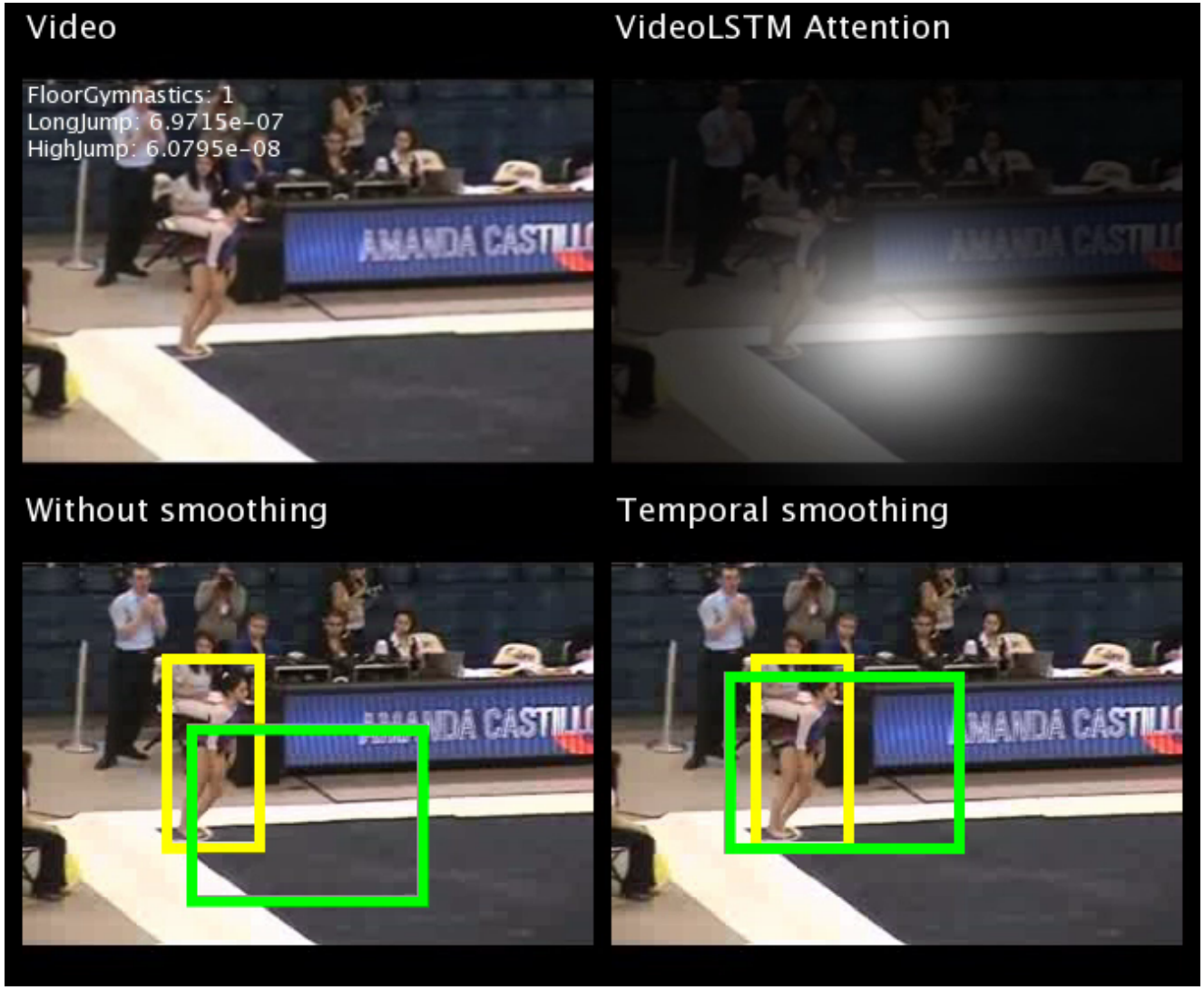
Thresholding + Connected components

Temporal smoothing



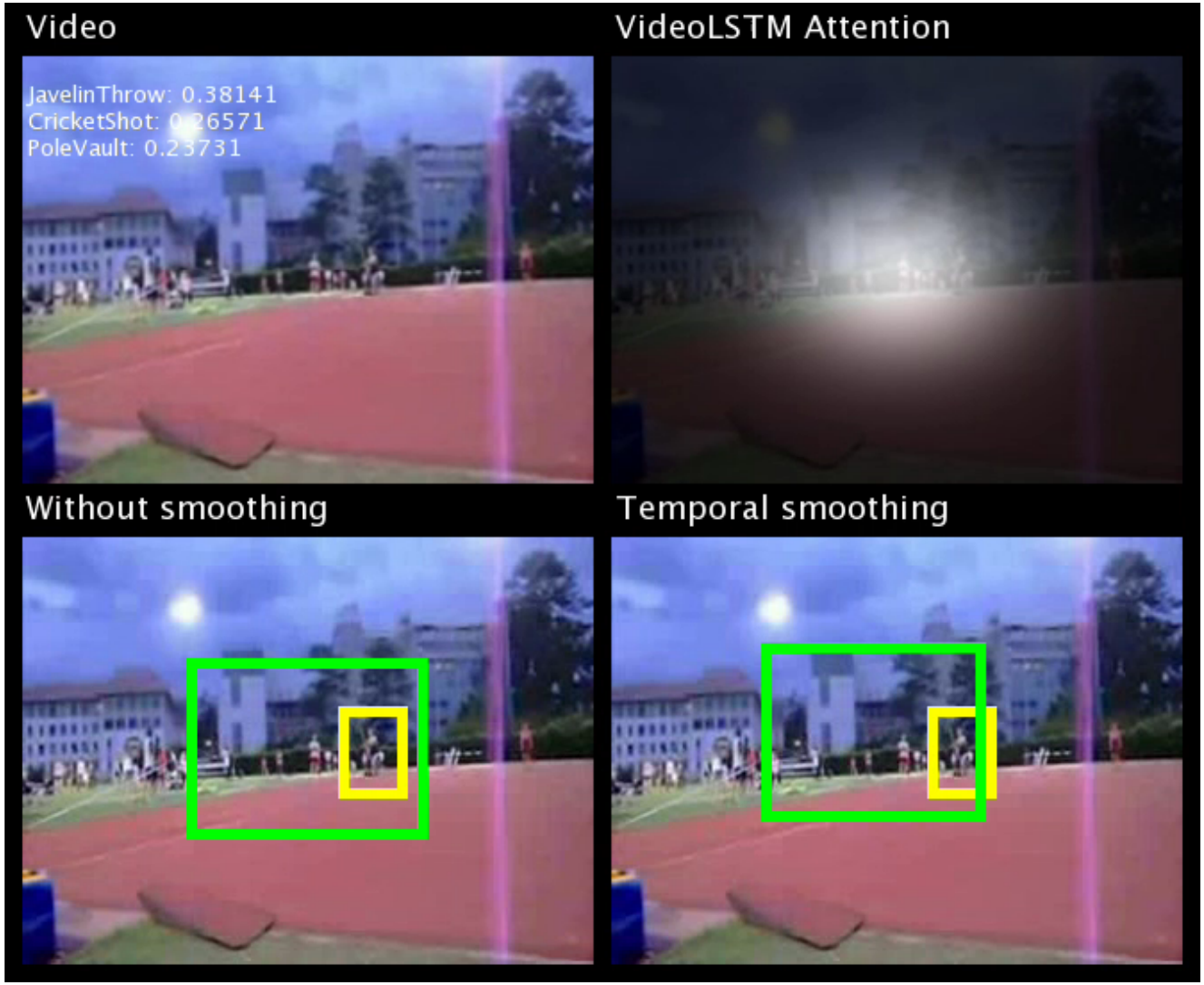
Action localization

Qualitative results



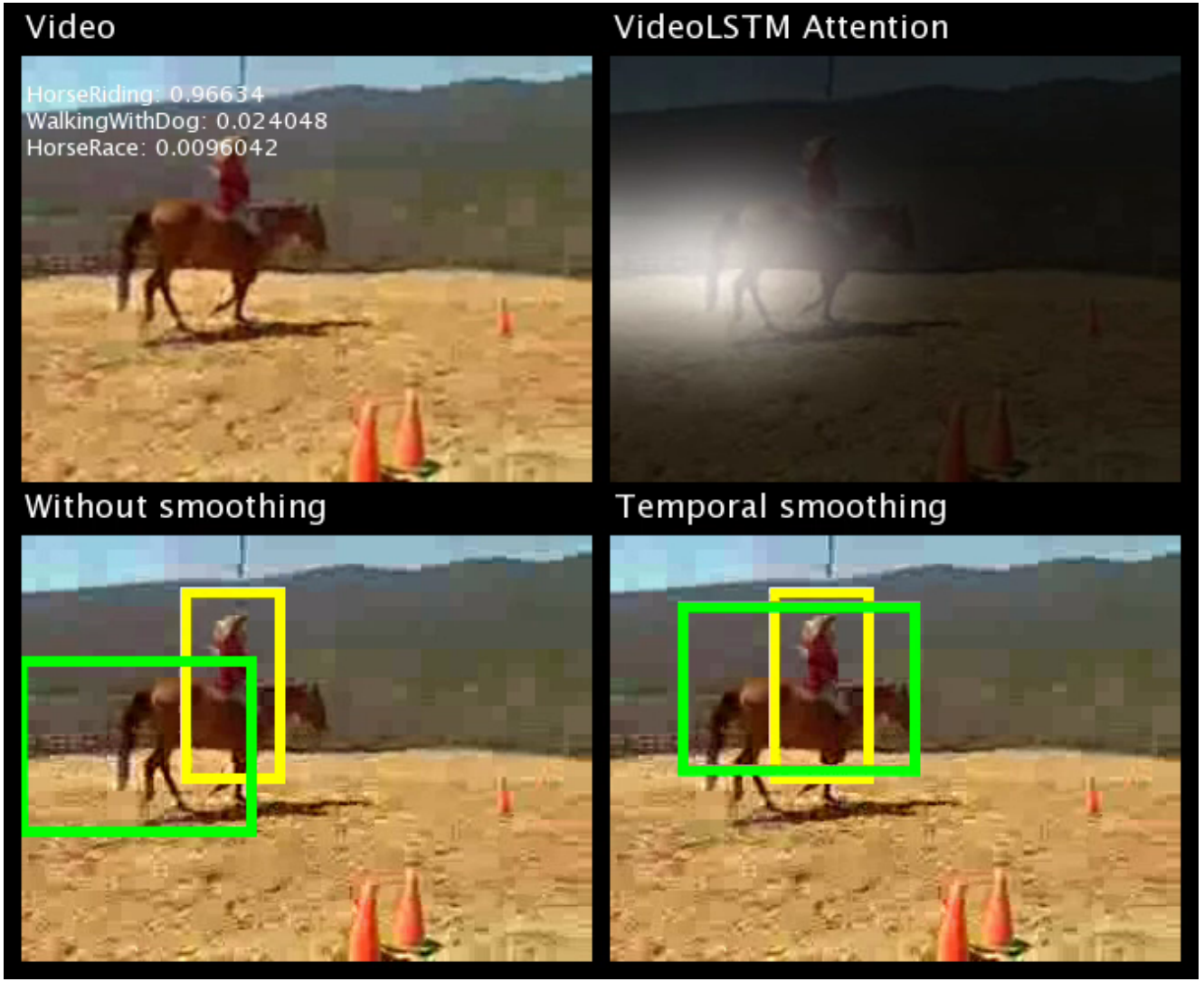
http://isis-data.science.uva.nl/zhenyang/videolstm/localization_examples/

Qualitative results



http://isis-data.science.uva.nl/zhenyang/videolstm/localization_examples/

Qualitative results



http://isis-data.science.uva.nl/zhenyang/videolstm/localization_examples/

Conclusions on VideoLSTM

Promising deep vision architecture for action localization

- Hardwires convolutions in attention LSTM
- Derives attention from what moves in video

Localization from a video-level action class label only

http://arxiv.org/abs/1607.01794