Activity Recognition

Computer Vision
Fall 2018
Columbia University

Many slides from Bolei Zhou

Project

- How are they going? About 30 teams have requested GPU credit so far
- Final presentations on December 5th and 10th
  - We will assign you to dates soon
- Final report due December 10 at midnight
- Details here: http://w4731.cs.columbia.edu/project

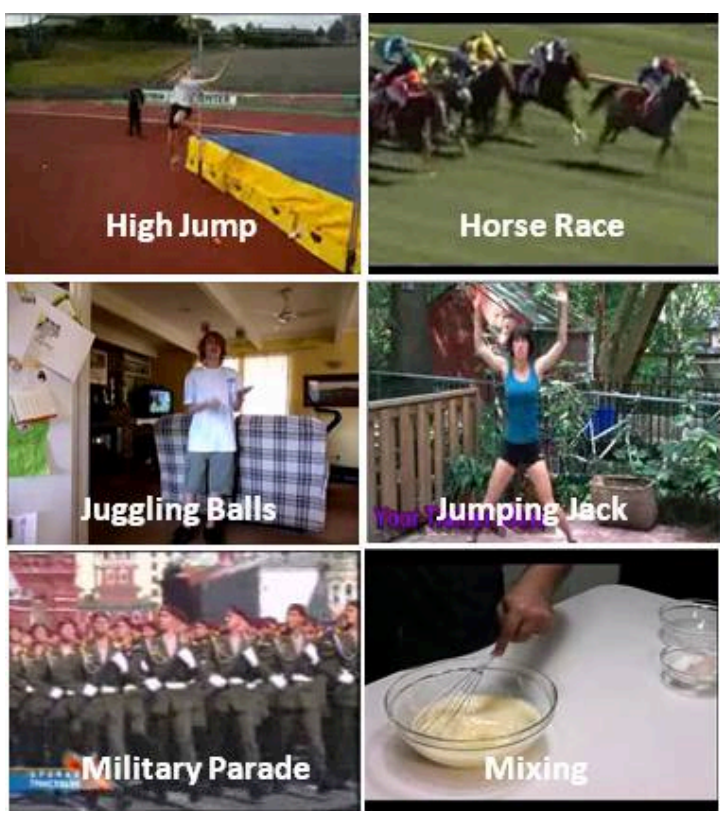
Challenge for Image Recognition

- Variation in appearance.

Challenge for Activity Recognition

- Describing activity at the proper level

Image recognition? Skeleton recognition?
No motion needed? Which activities?



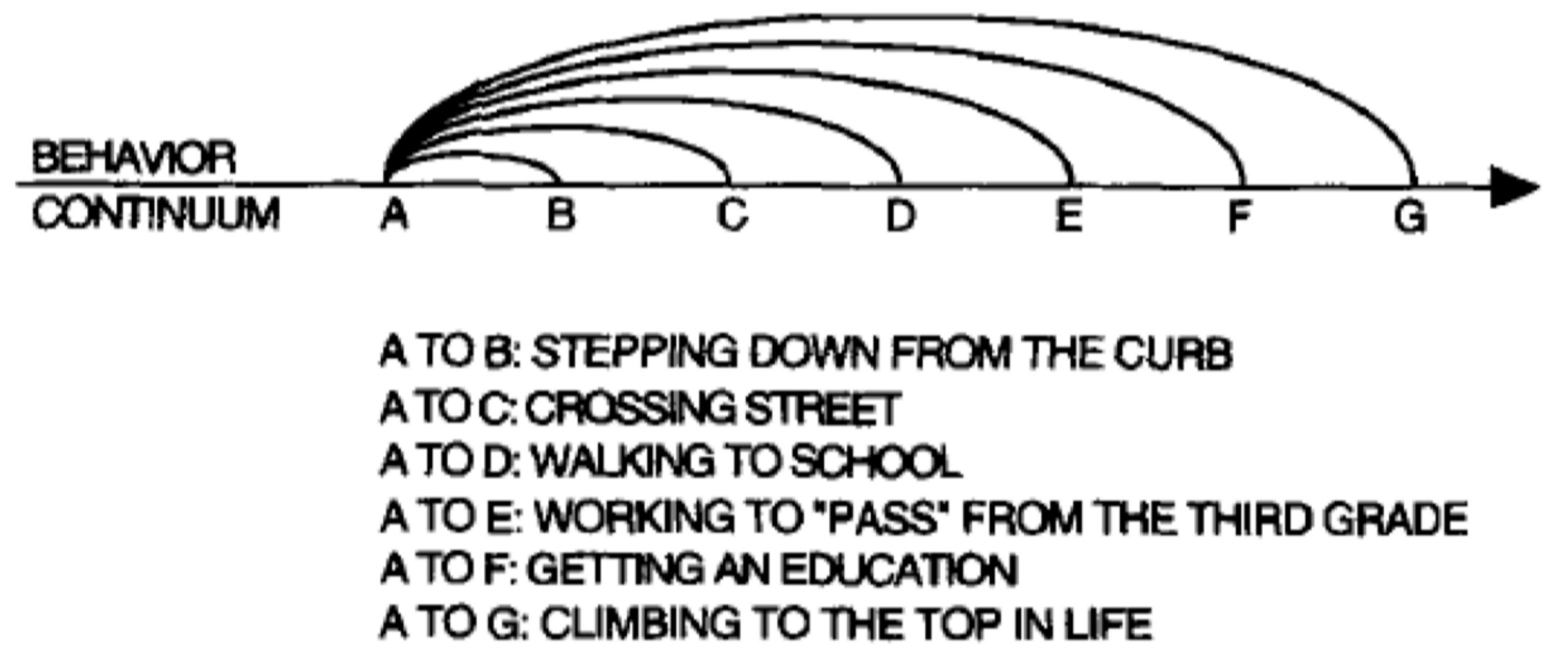
Challenge for Activity Recognition

- Describing activity at the proper level

A chain of events
Making chocolate cookies

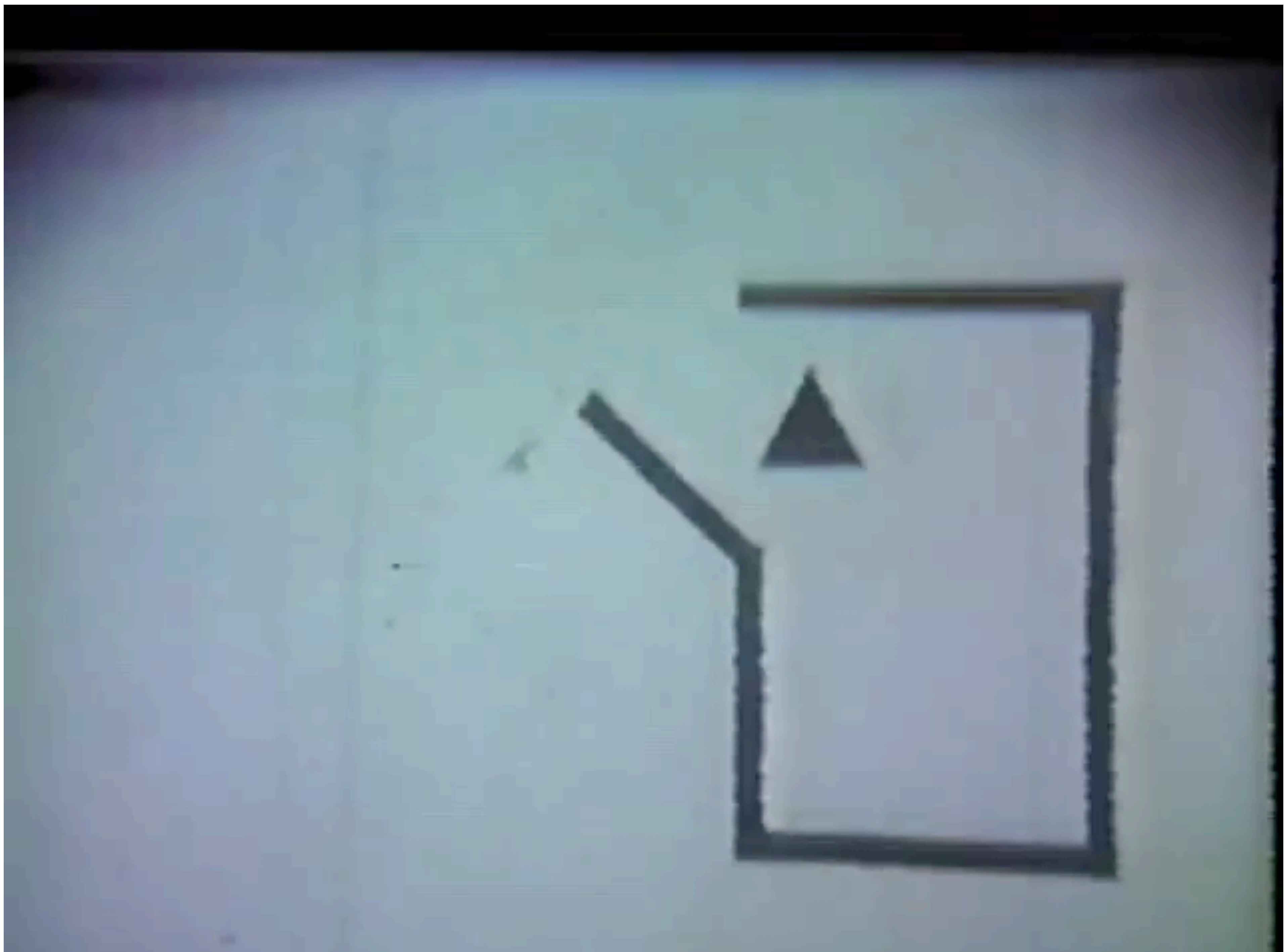
What are they doing?

What are they doing?



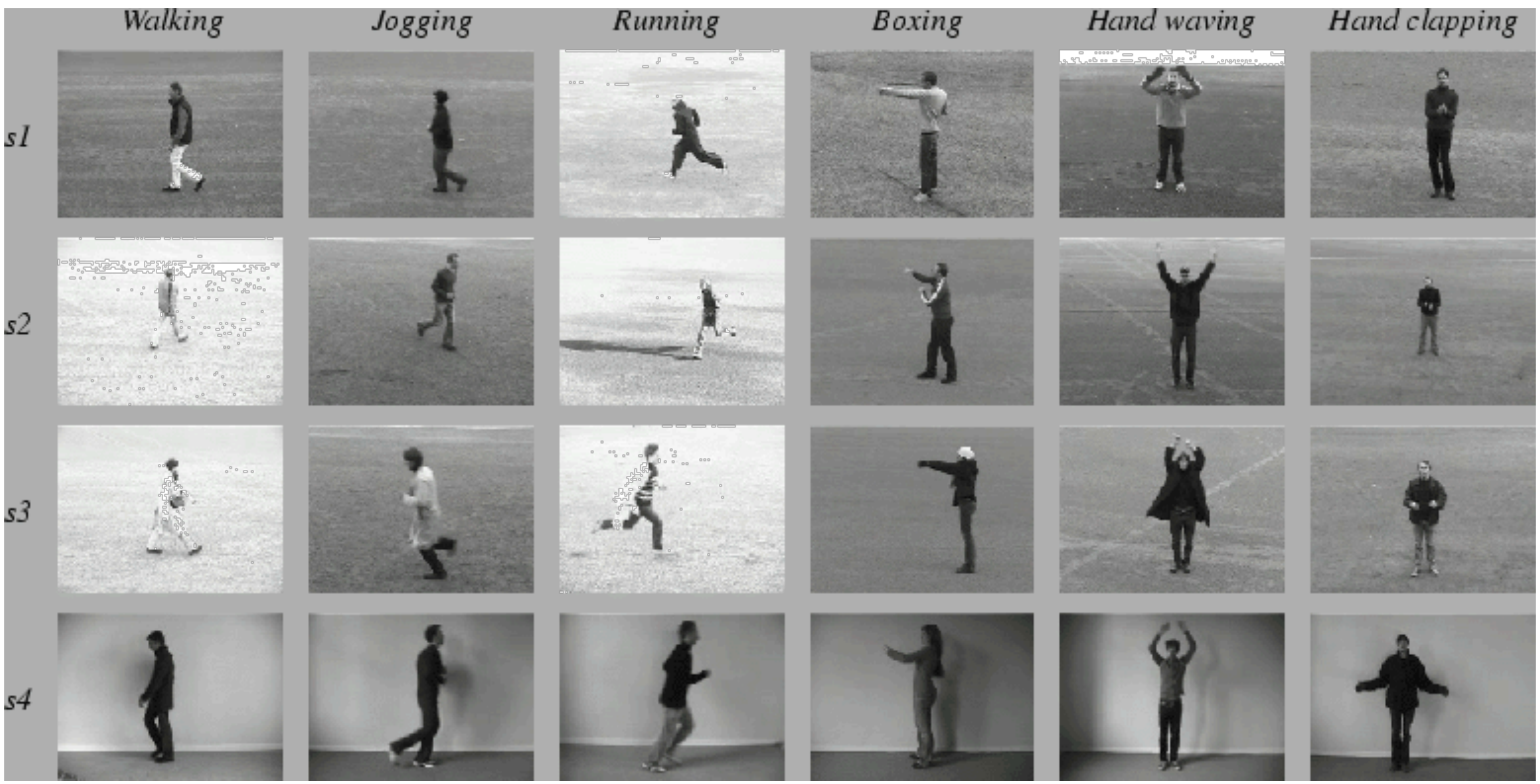
Barker and Wright, 1954

Vision or Cognition?



Video Recognition Datasets

- KTH Dataset: recognition of human actions
- 6 classes, 2391 videos

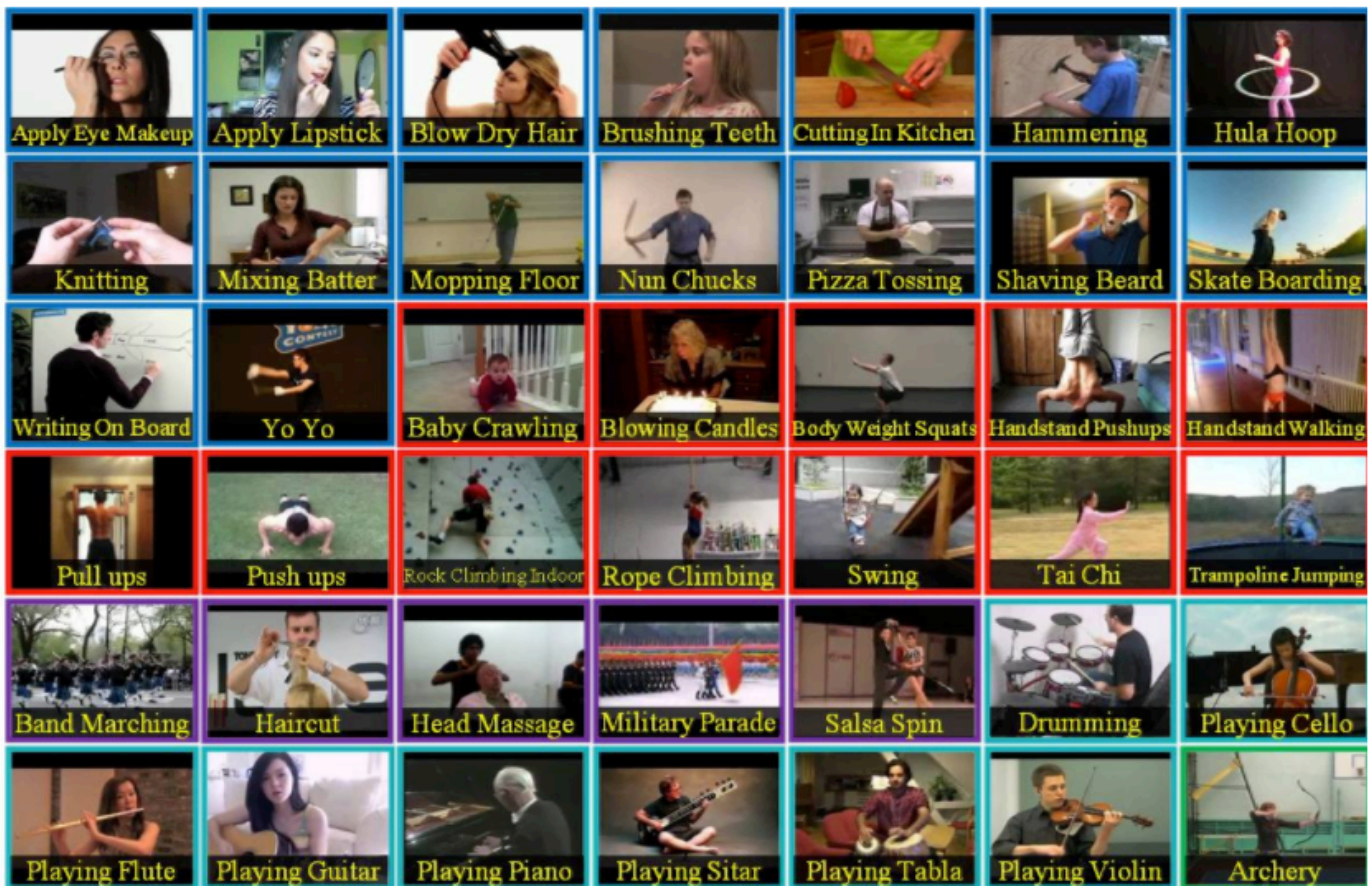


https://www.youtube.com/watch?v=Jm69kbCC17s

Recognizing Human Actions: A Local SVM Approach. ICPR 2004

Video Recognition Datasets

- UCF101 from University of Central Florida
- 101 classes, 9,511 videos in training

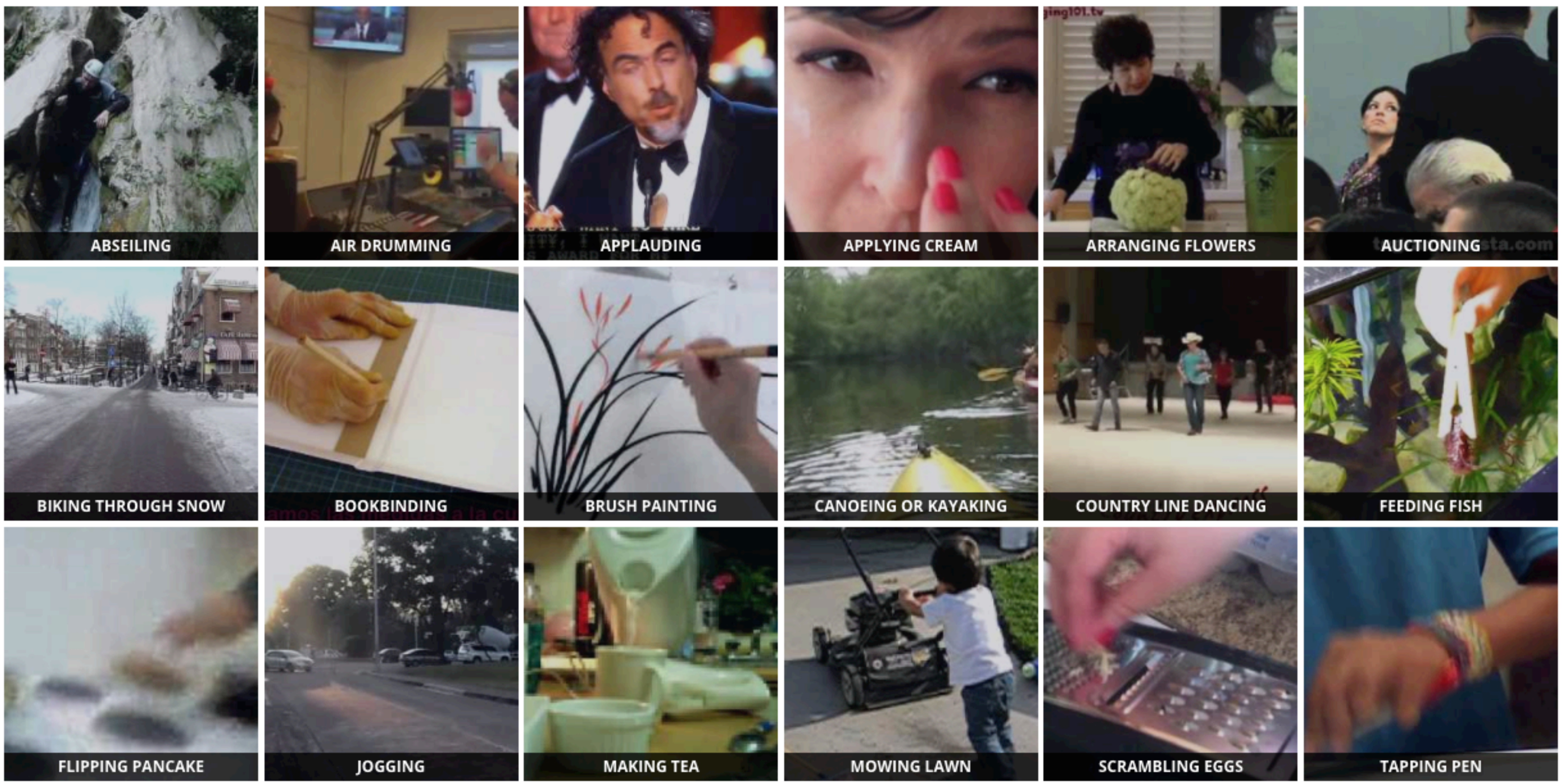


https://www.youtube.com/watch?v=hGhuUaxocIE

UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. 2012

Video Recognition Datasets

- Kinetics from Google DeepMind
- 400 classes, 239,956 videos in training



https://deepmind.com/research/open-source/open-source-datasets/kinetics/

Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowdsourced video dataset

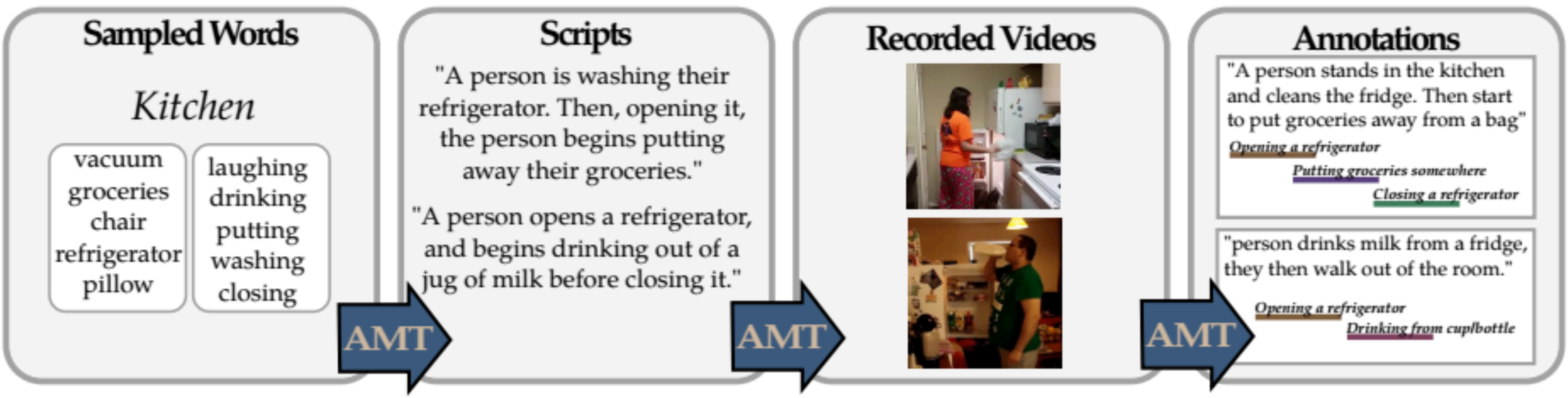
http://allenai.org/plato/charades/

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV'16

Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowdsourced video dataset

amazon mechanical turk



Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV'16

Example annotated videos from the Charades dataset

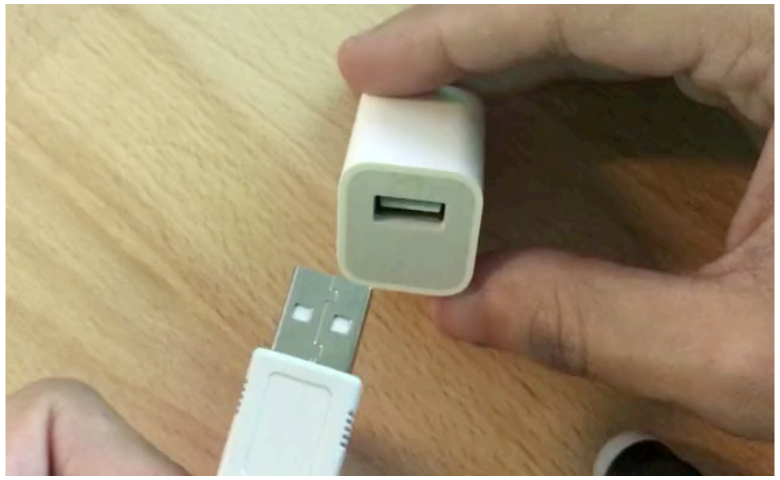
Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding

Video Recognition Datasets

- Something-Something dataset: human object interaction
- 174 categories: 100,000 videos

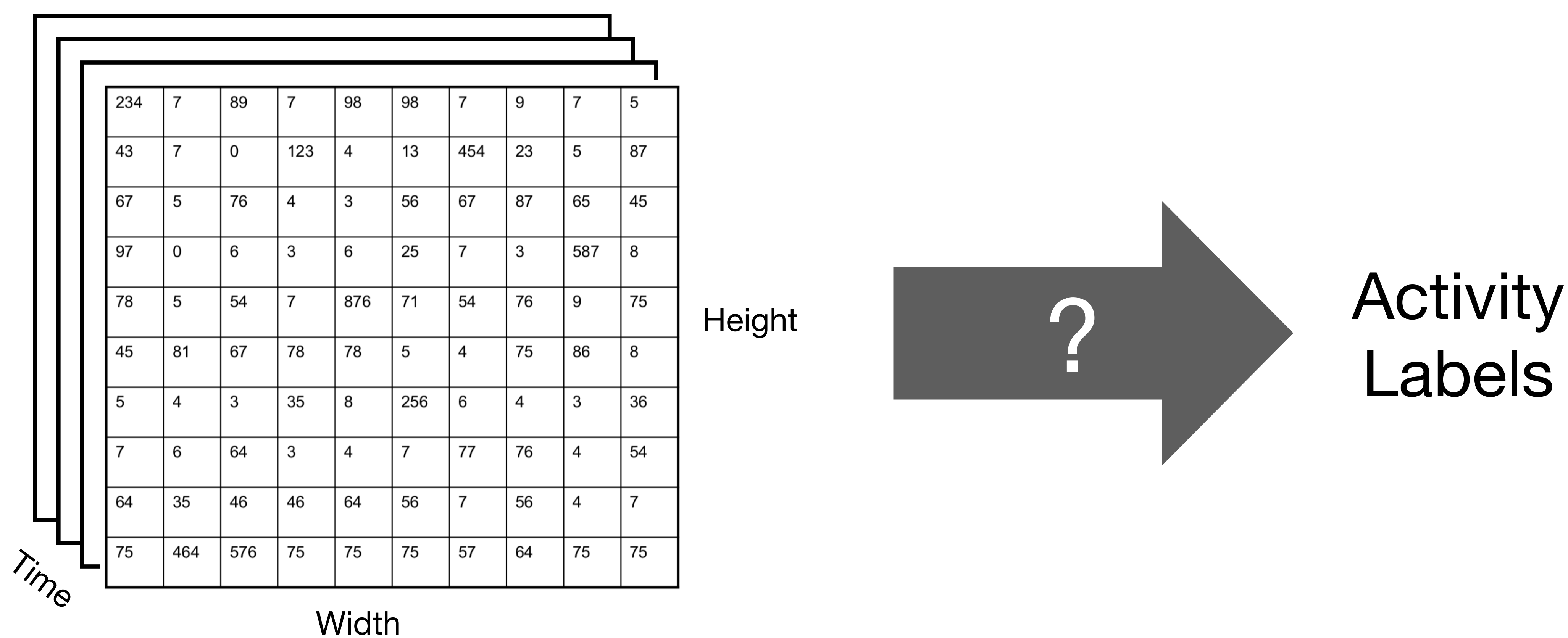
- Holding something
- Turning something upside down
- Turning the camera left while filming something
- Opening something

Poking a stack of something so the stack collapses



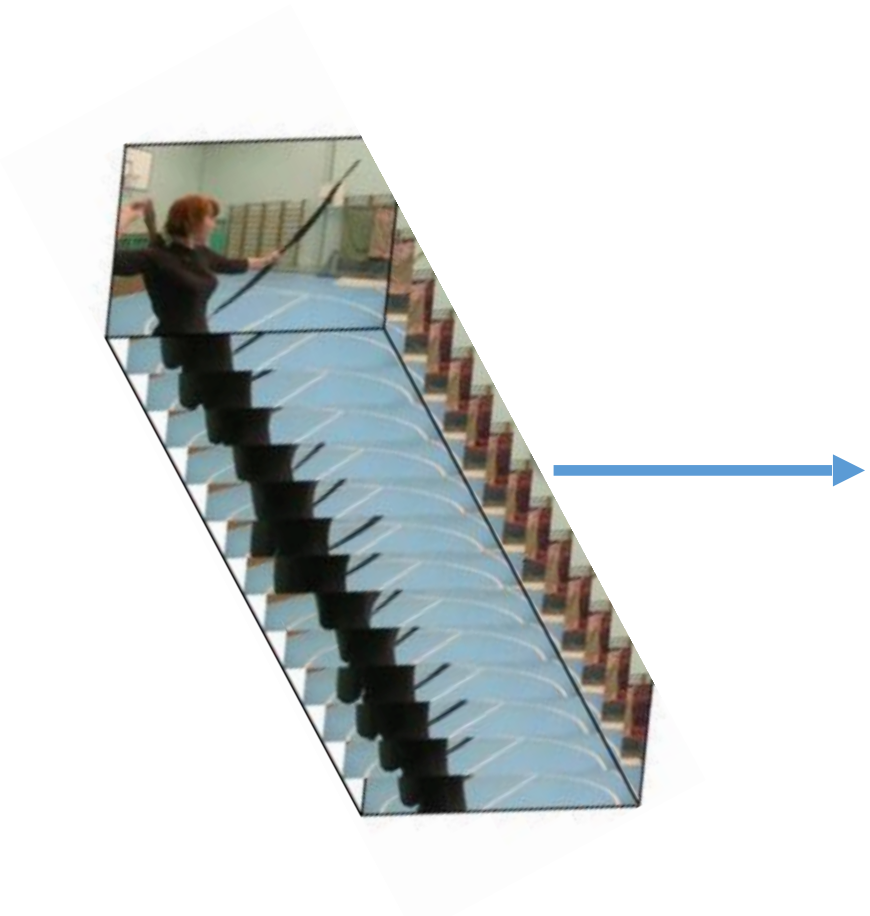
Plugging something into something

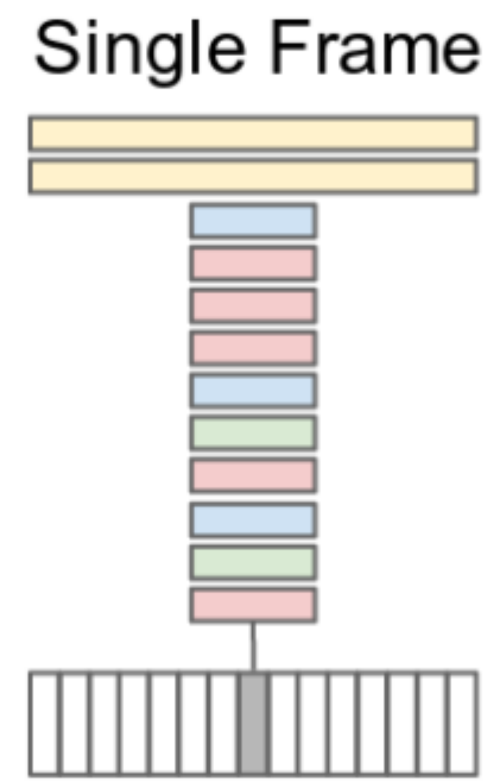
https://www.twentybn.com/datasets/something-something



234	7	89	7	98	98	7	9	7	5
43	7	0	123	4	13	454	23	5	87
67	5	76	4	3	56	67	87	65	45
97	0	6	3	6	25	7	3	587	8
78	5	54	7	876	71	54	76	9	75
45	81	67	78	78	5	4	75	86	8
5	4	3	35	8	256	6	4	3	36
7	6	64	3	4	7	77	76	4	54
64	35	46	46	64	56	7	56	4	7
75	464	576	75	75	75	57	64	75	75

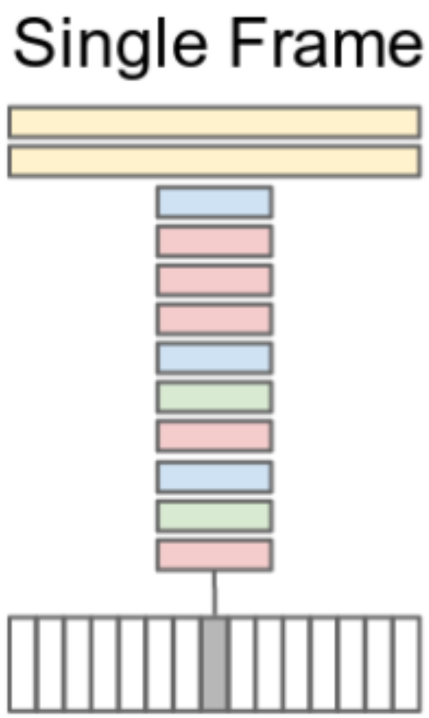
Single-frame image model





Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

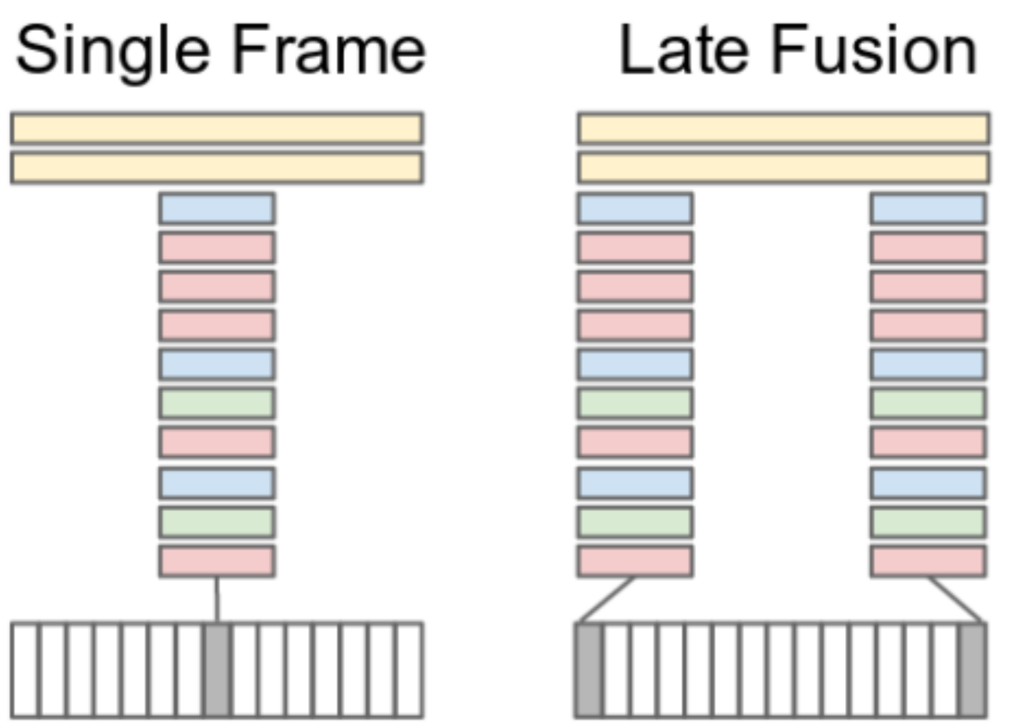
Multi-frame fusion model



41.1%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

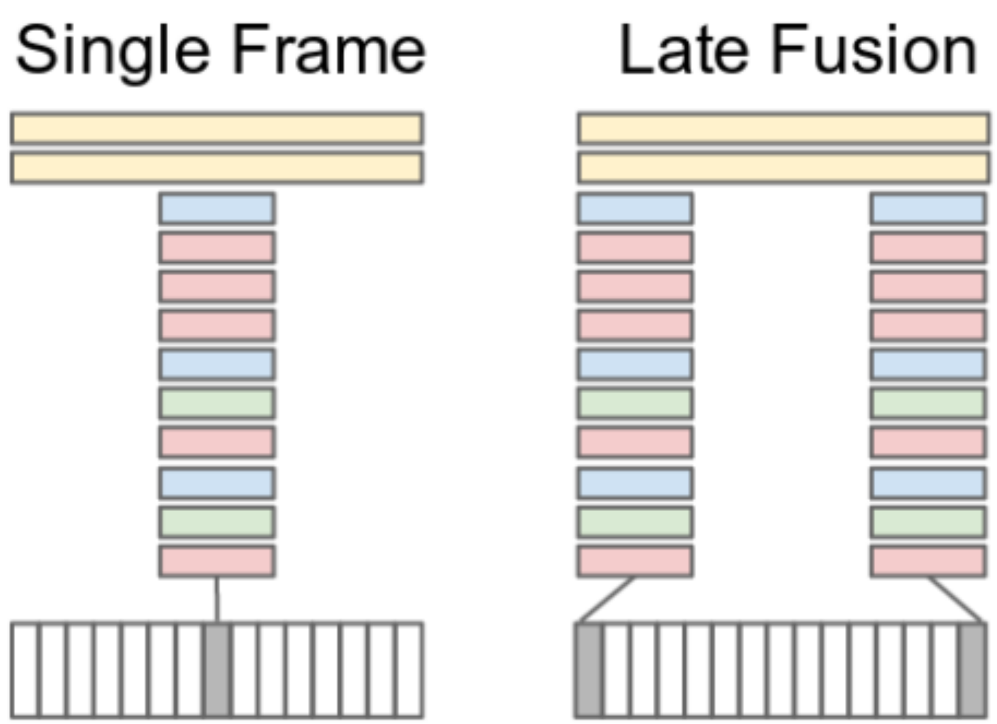
Multi-frame fusion model



41.1%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

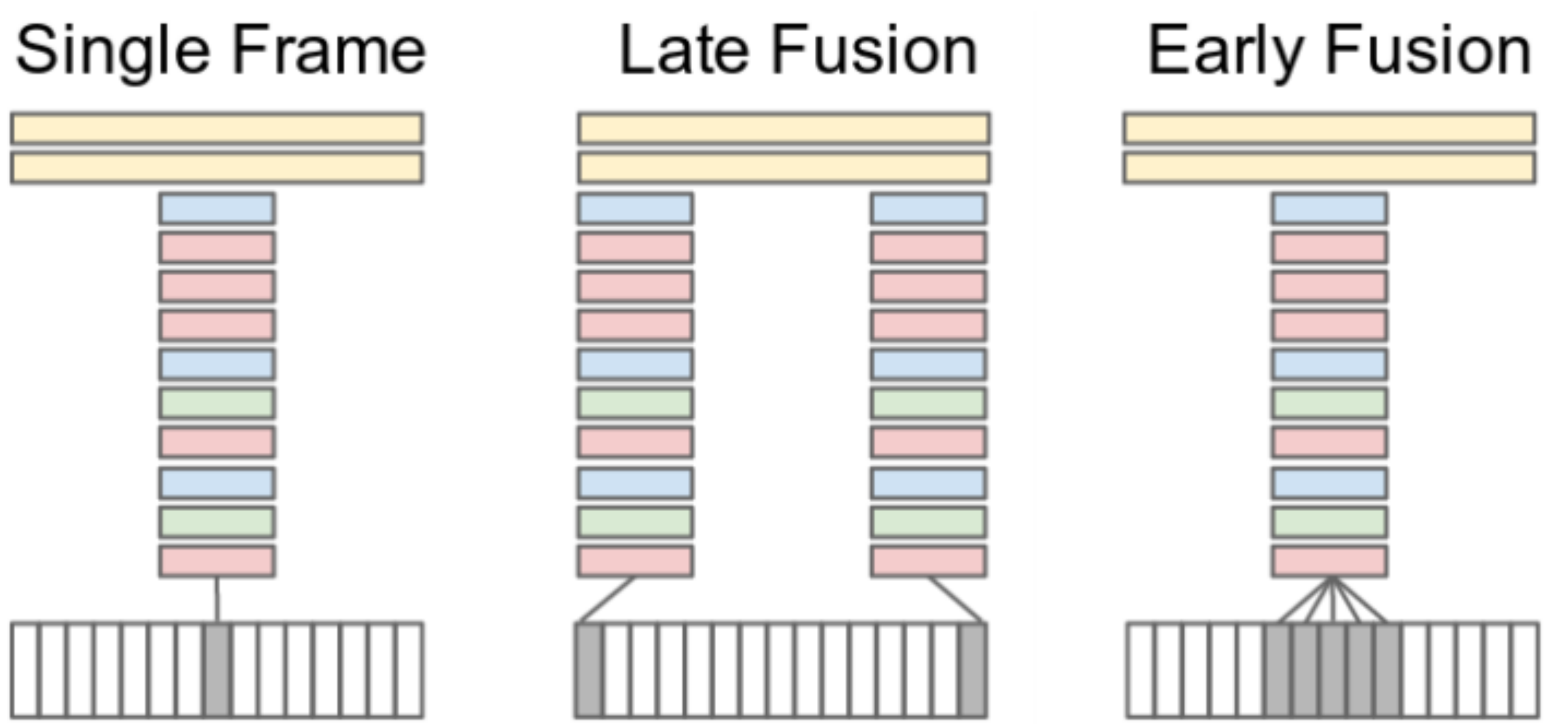
Multi-frame fusion model



41.1% 40.7%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

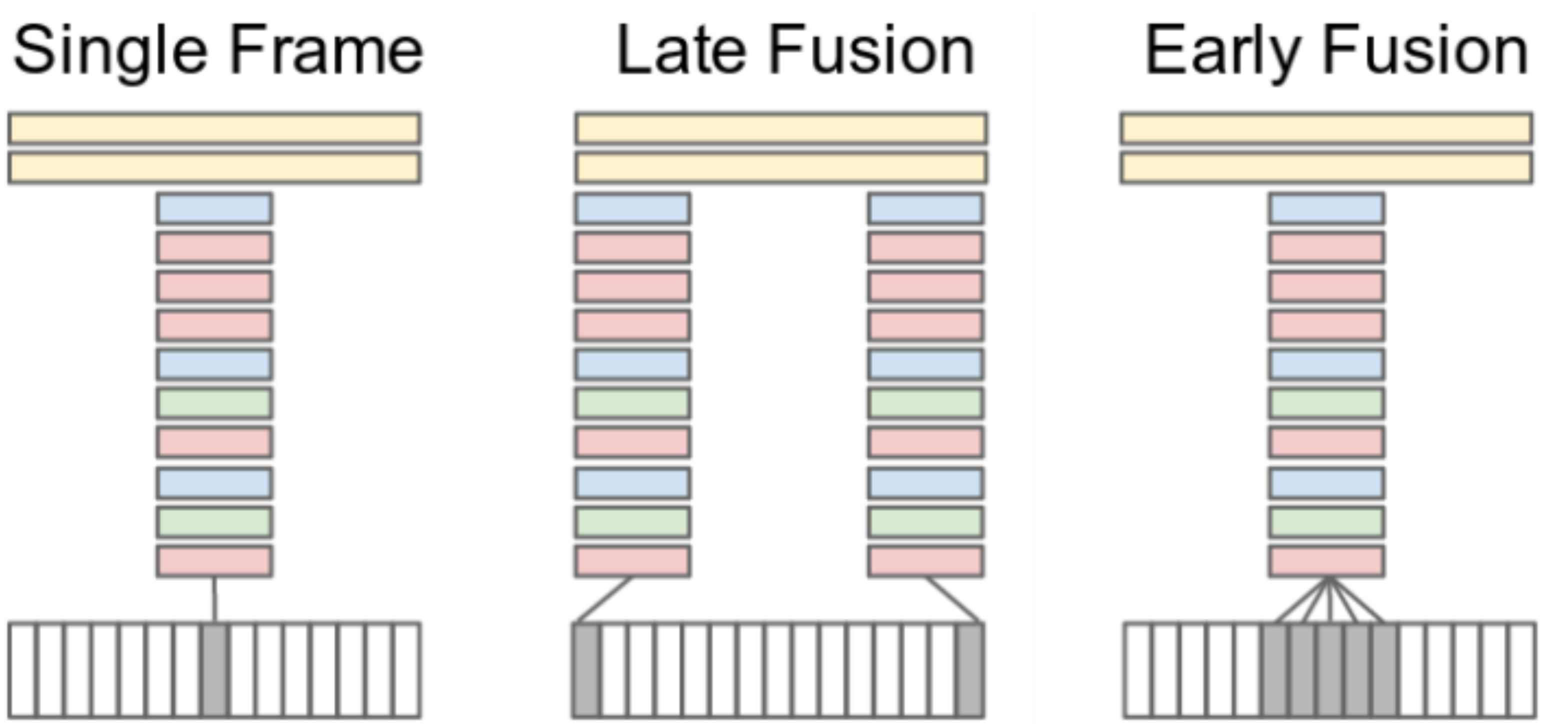
Multi-frame fusion model



41.1% 40.7%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

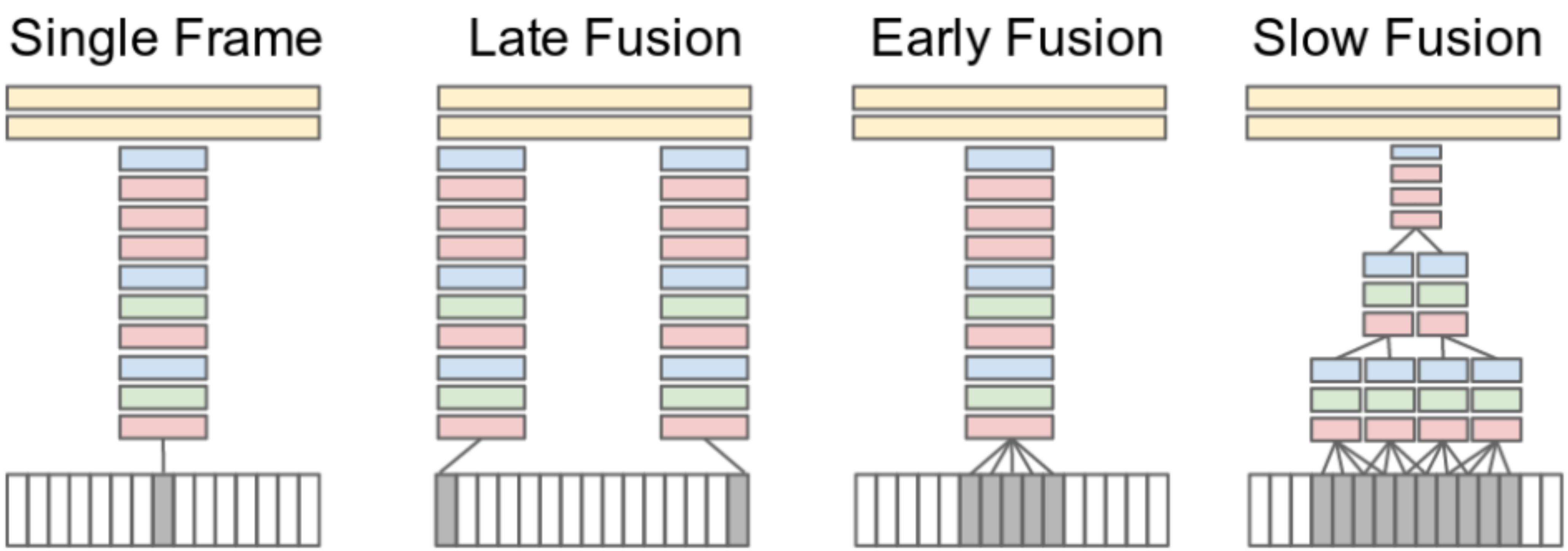
Multi-frame fusion model



41.1% 40.7% 38.9%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

Multi-frame fusion model

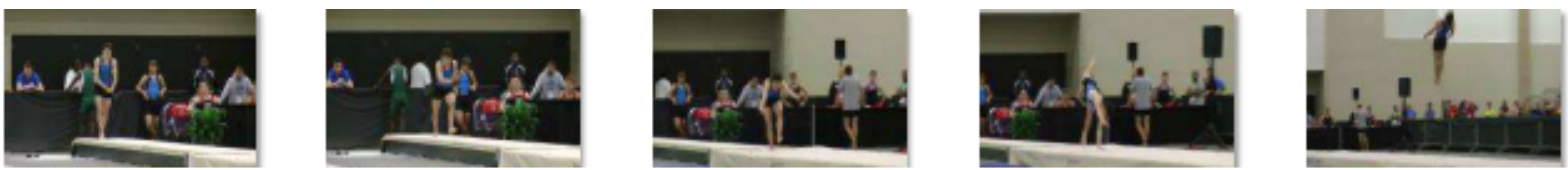


41.1% 40.7% 38.9% 41.9%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

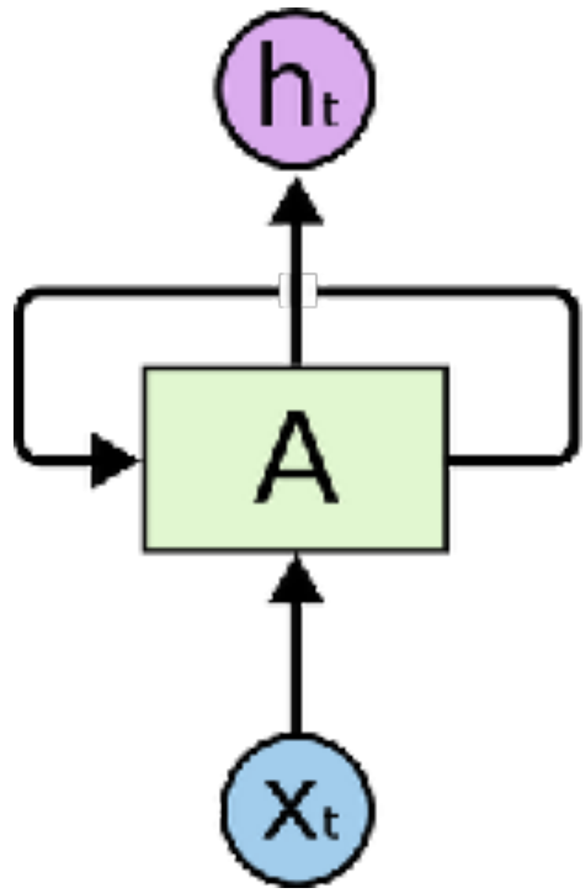
Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

Sequence of frames?



Long-term Recurrent Convolutional Networks for Visual Recognition and Description.
CVPR 2015

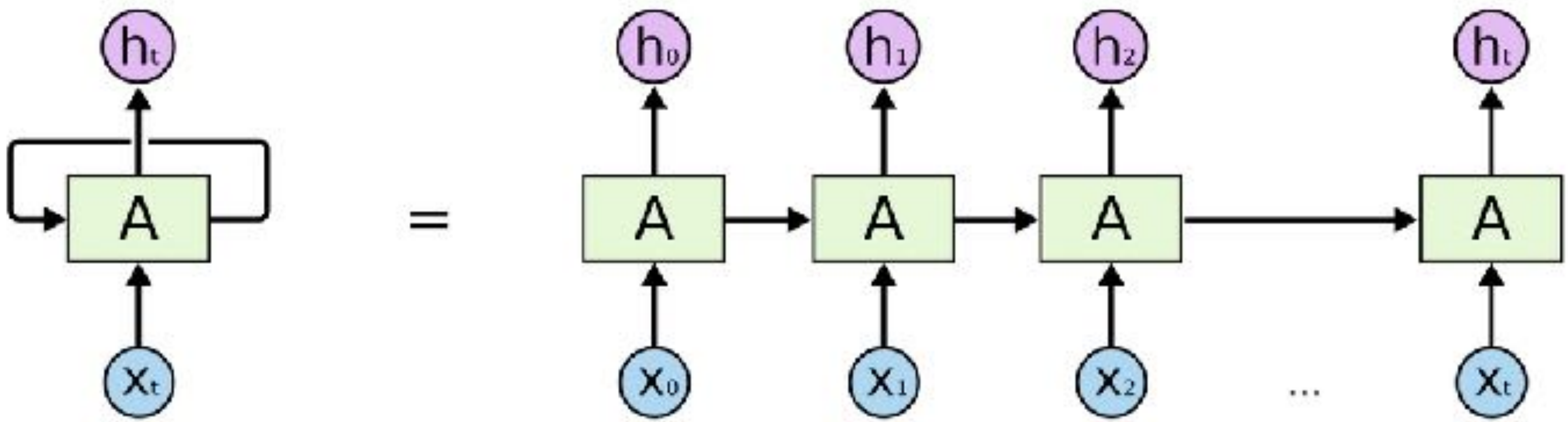
Recurrent Neural Networks (RNNs)



In the above diagram, a chunk of neural network, A, looks at some input x_i and outputs a value h_i. A loop allows information to be passed from one step of the network to the next.

Credit: Christopher Olah

Recurrent Neural Networks (RNNs)

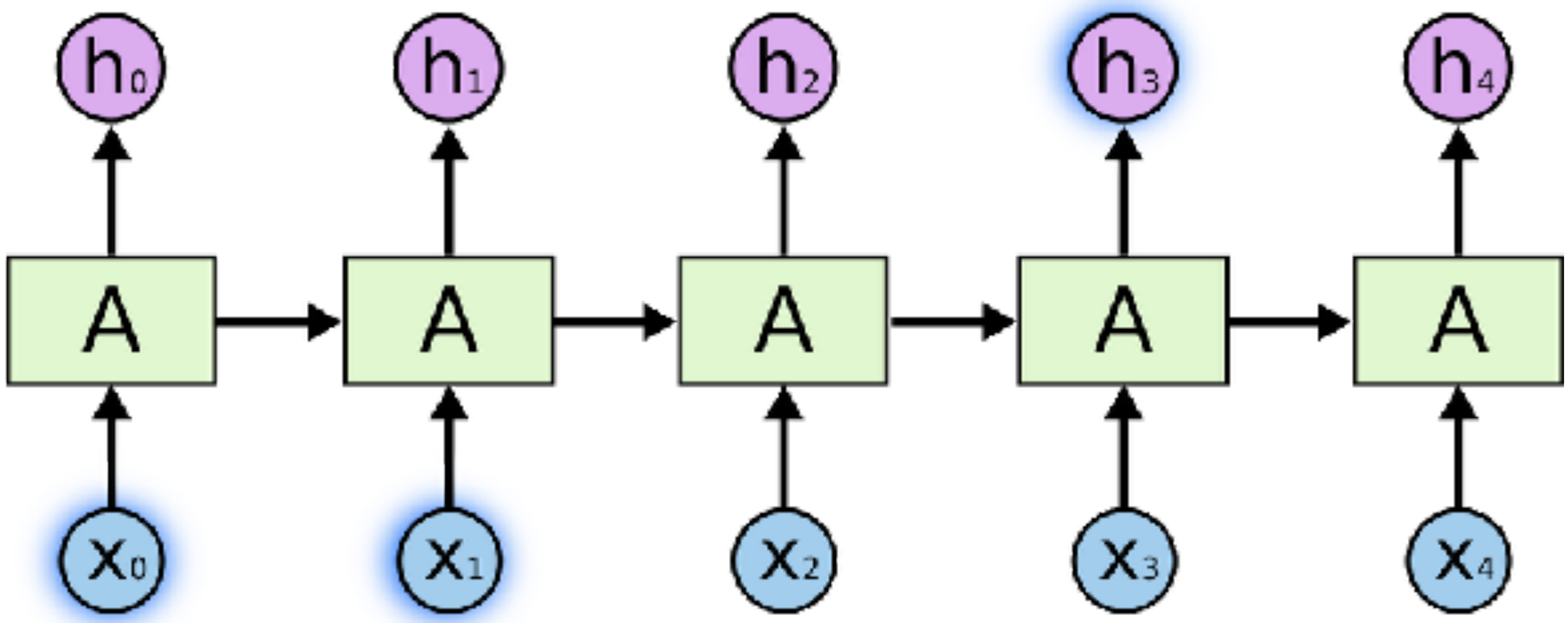


An unrolled recurrent neural network.

A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor

Credit: Christopher Olah

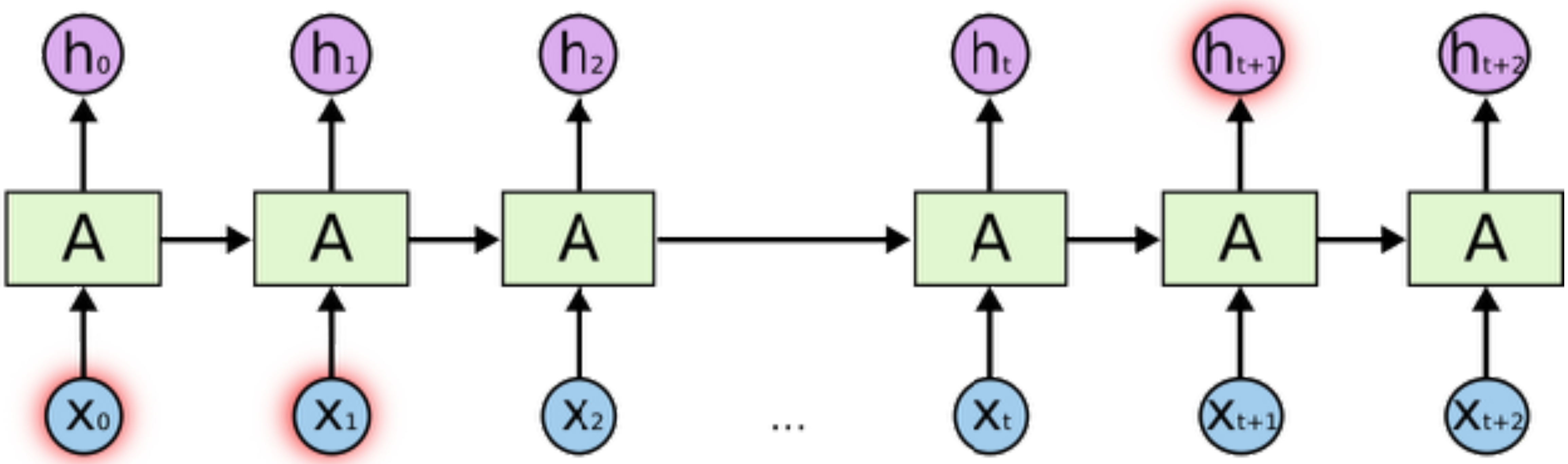
Recurrent Neural Networks (RNNs)



When the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information

Credit: Christopher Olah

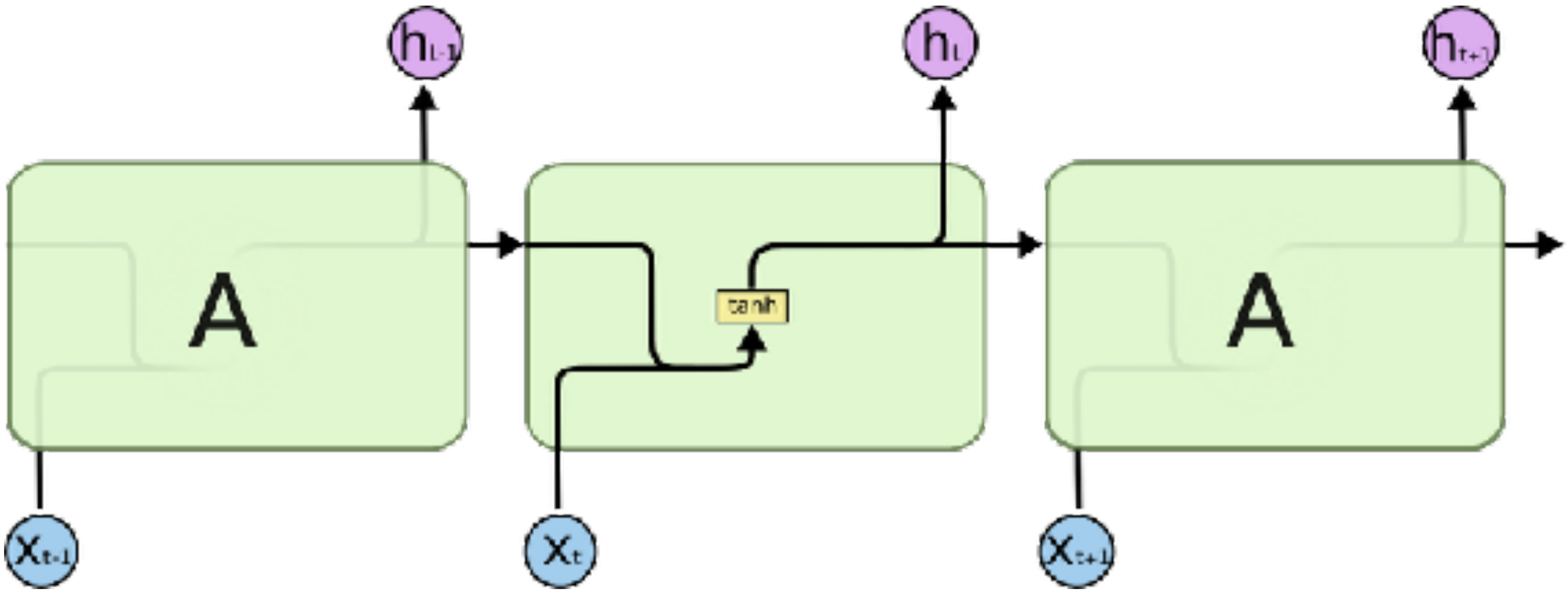
Long-term dependencies - hard to model!



But there are also cases where we need more context.

Credit: Christopher Olah

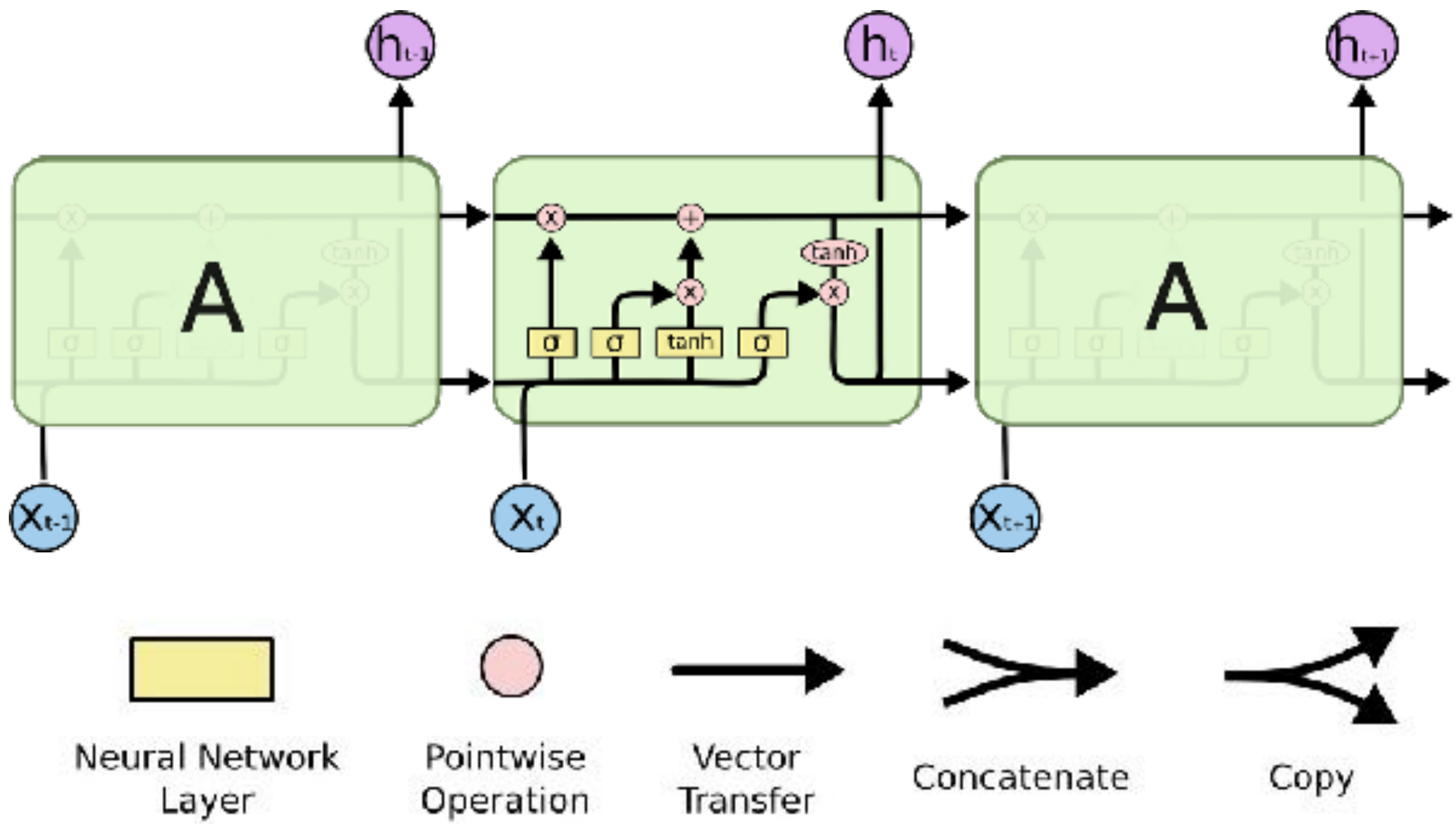
From plain RNNs to LSTMs



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Credit: Christopher Olah

From plain RNNs to LSTMs

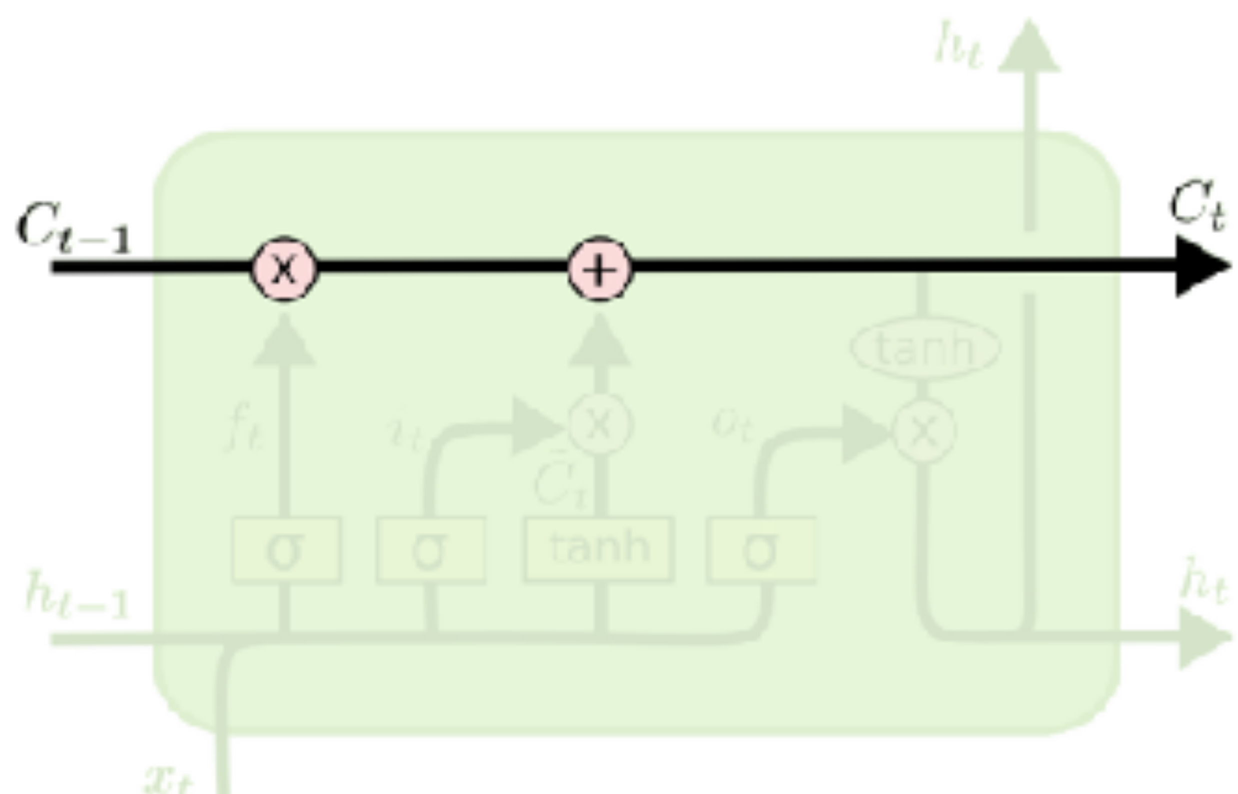


(LSTM: Long Short Term Memory Networks)

Credit: Christopher Olah

LSTMs Step by Step: Memory

Cell State / Memory

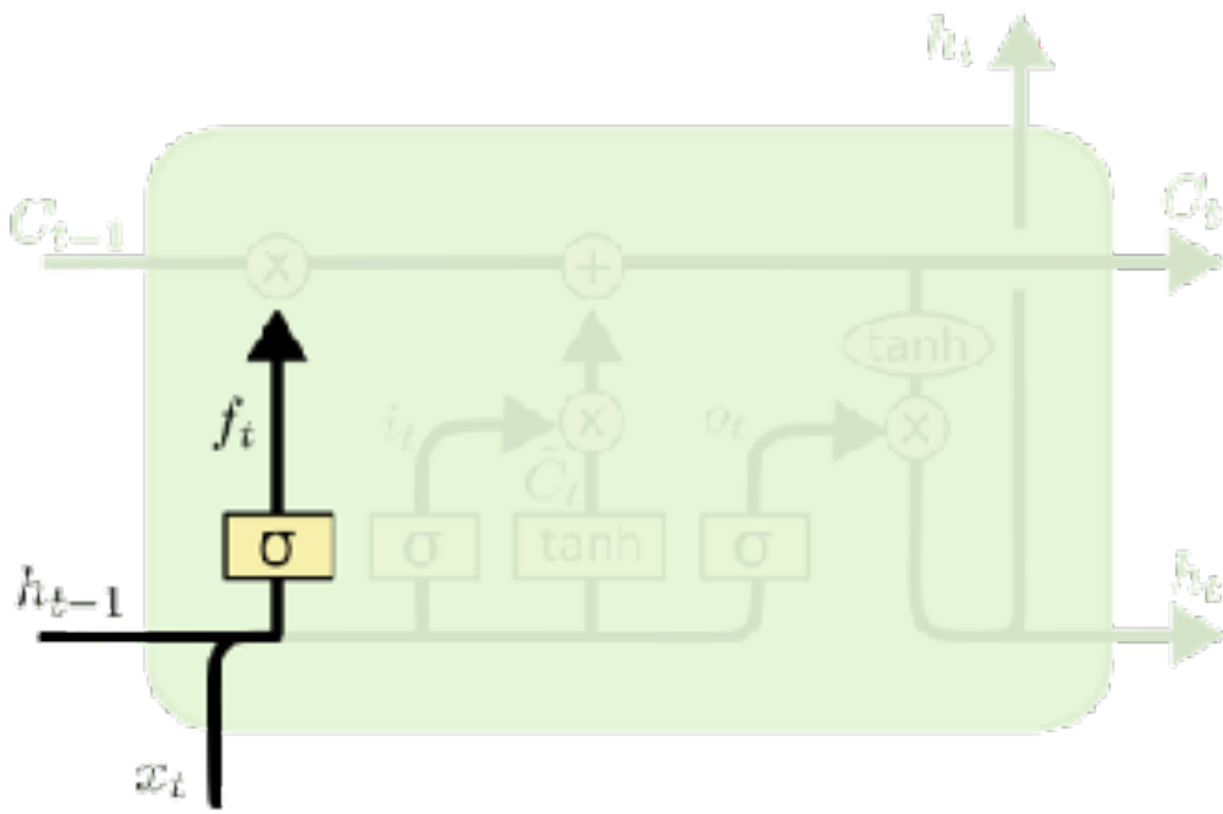


The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates

Credit: Christopher Olah

LSTMs Step by Step: Forget Gate

Should we continue to remember this "bit" of information or not?



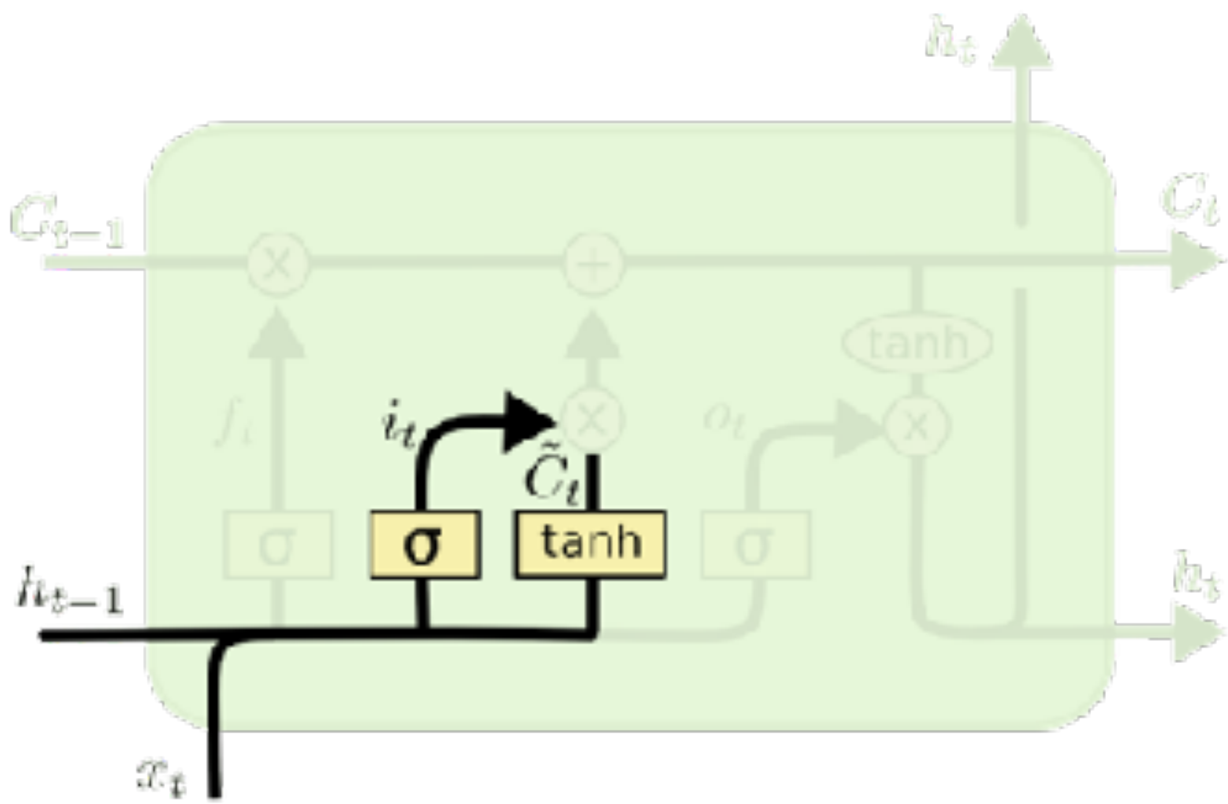
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

The first step in our LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer."

Credit: Christopher Olah

LSTMs Step by Step: Input Gate

Should we update this “bit” of information or not? If so, with what?



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

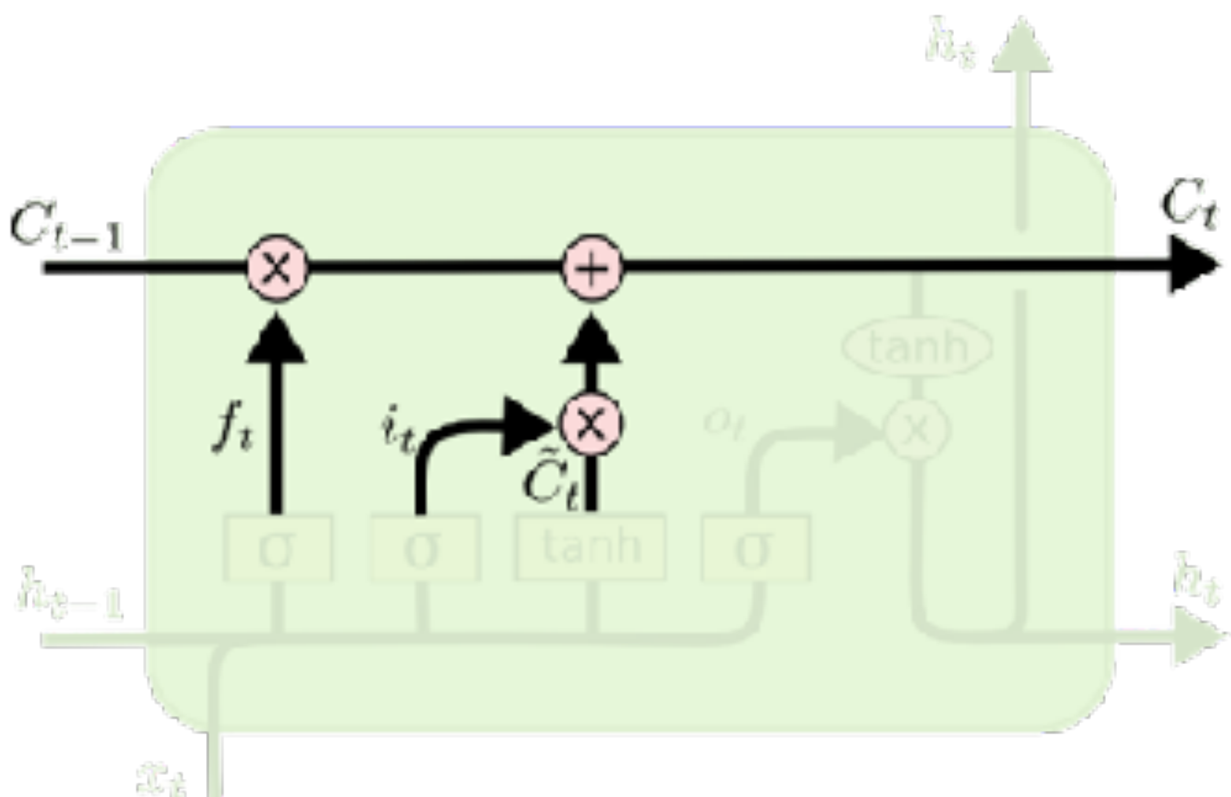
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

The next step is to decide what new information we’re going to store in the cell state. This has two parts. First, a sigmoid layer called the “input gate layer” decides which values we’ll update. Next, a tanh layer creates a vector of new candidate values, $\tilde{C}_t$, that could be added to the state.

Credit: Christopher Olah

LSTMs Step by Step: Memory Update

Decide what will be kept in the cell state/memory

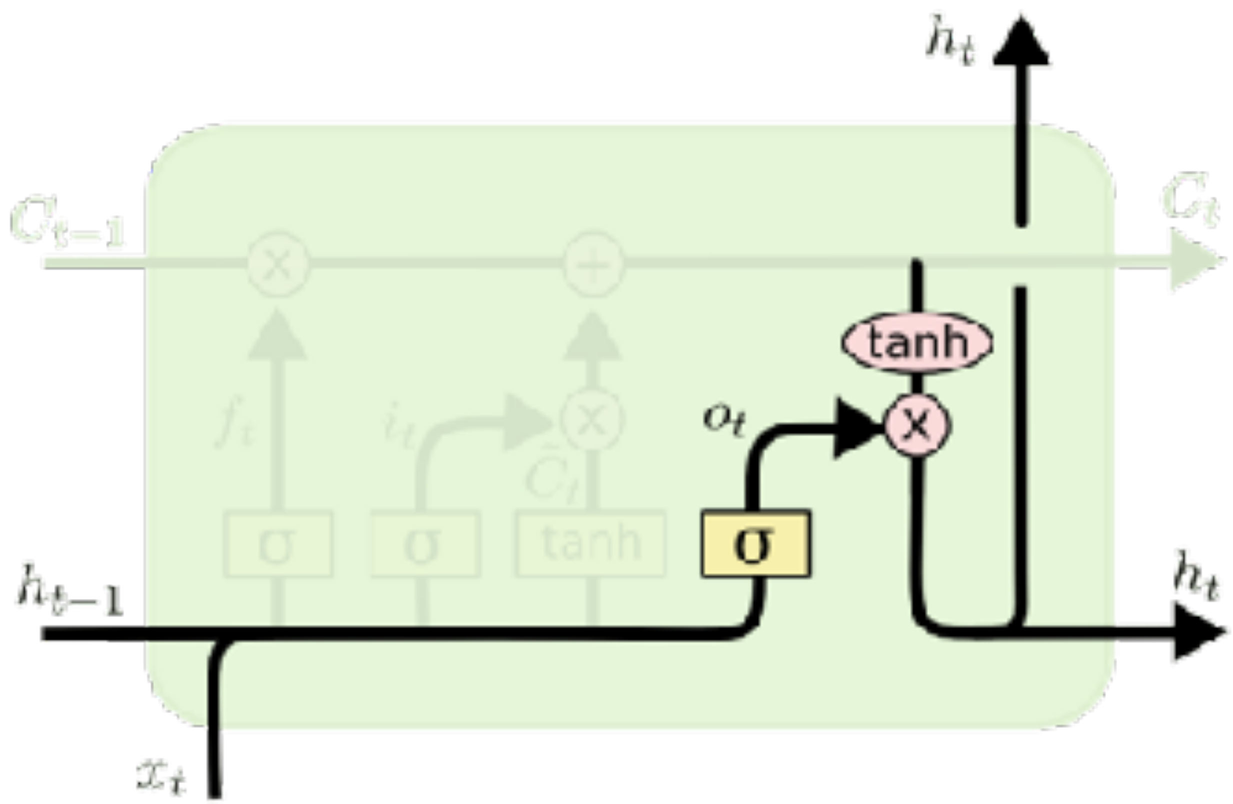


$$C_t = \underbrace{f_t * C_{t-1}}_{\text{Forget that}} + \underbrace{i_t * \tilde{C}_t}_{\text{Memorize this}}$$

Credit: Christopher Olah

LSTMs Step by Step: Output Gate

Should we output this "bit" of information?



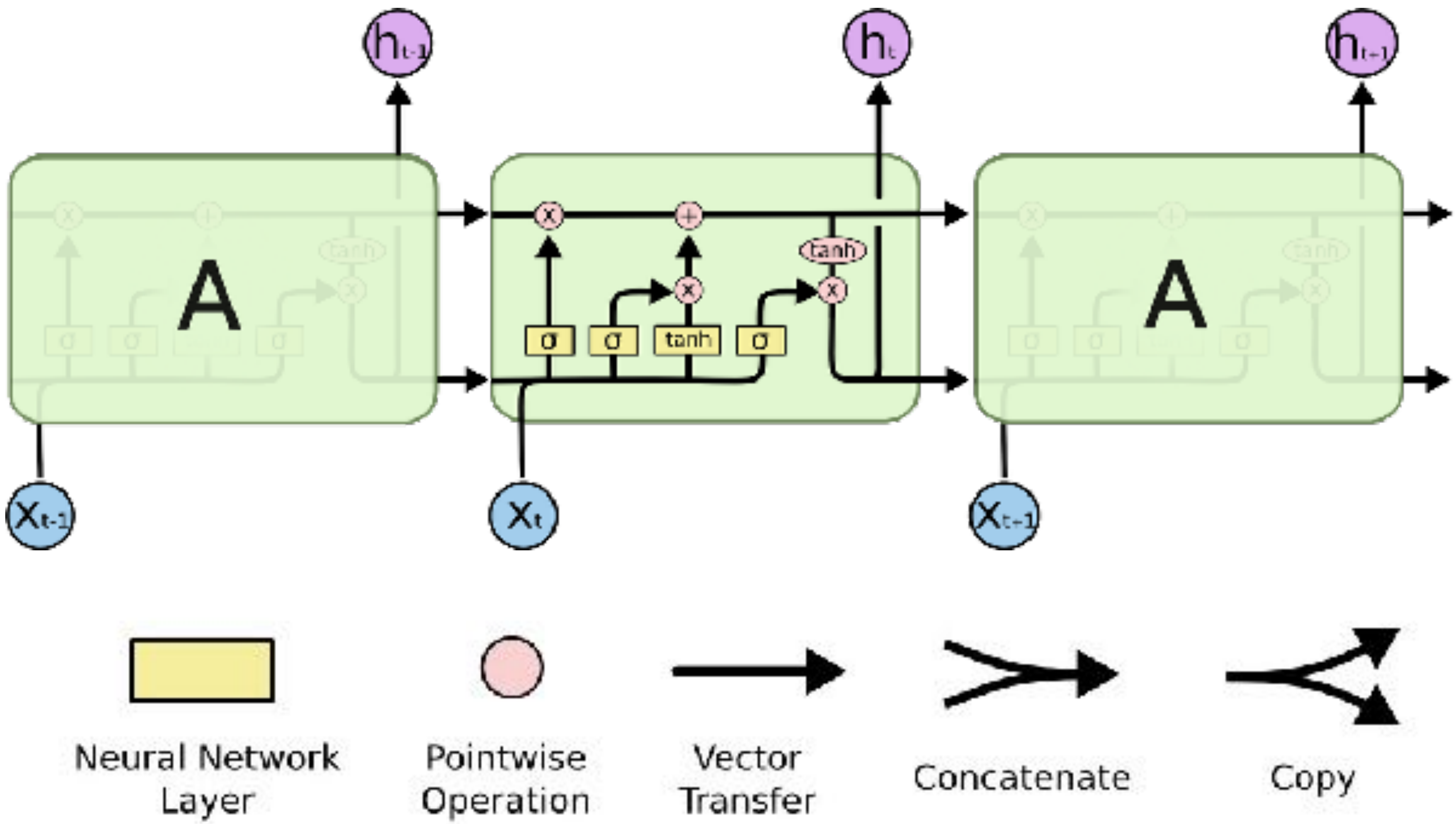
$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we're going to output. Then, we put the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

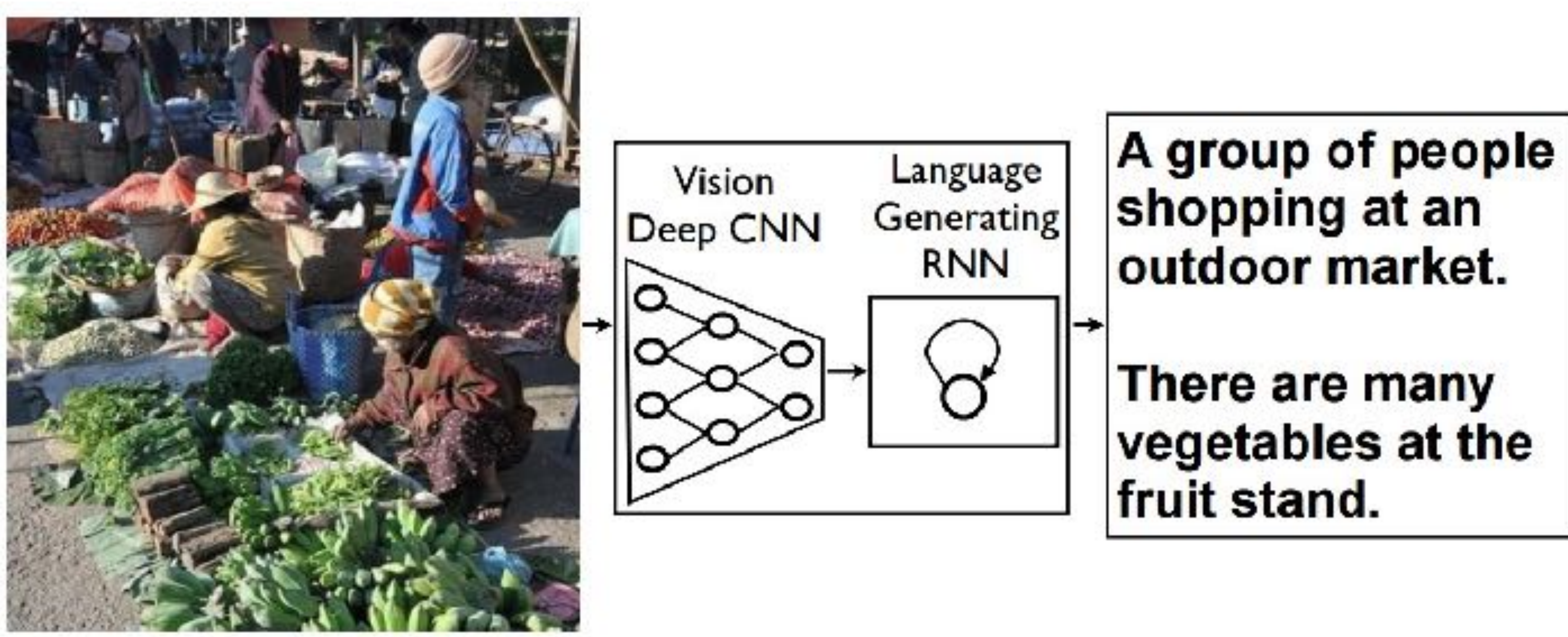
Credit: Christopher Olah

Complete LSTM - A pretty sophisticated cell



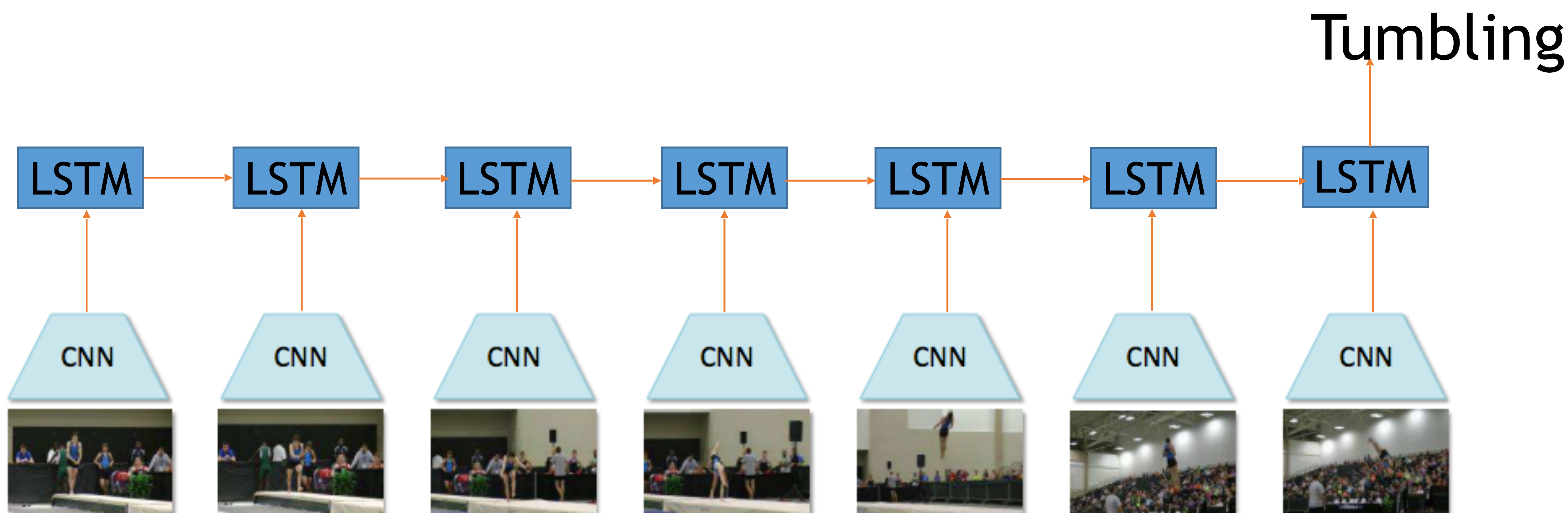
Credit: Christopher Olah

Show and Tell: A Neural Image Caption Generator



Show and Tell: A Neural Image Caption Generator, Vinyals et. al., CVPR 2015

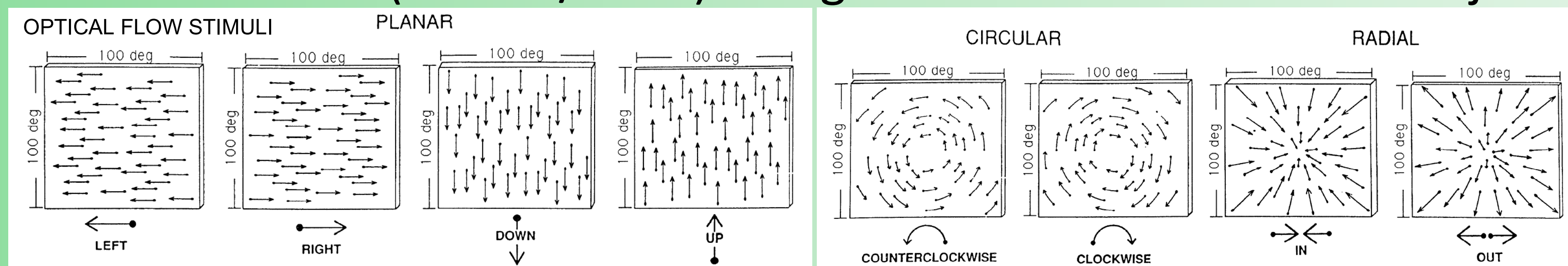
Multi-frame LSTM fusion model

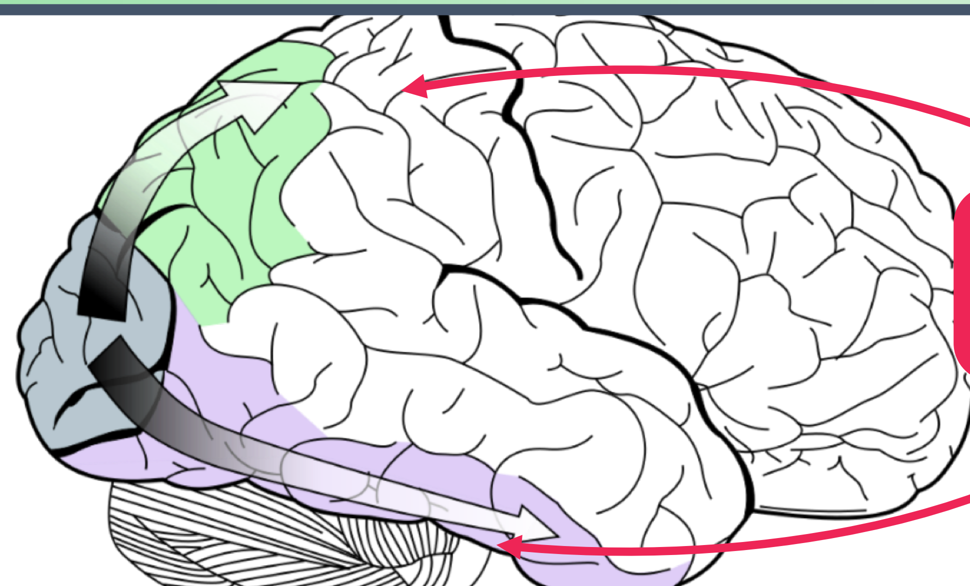


Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015

Motivation: Separate visual pathways in nature

➔ Dorsal stream (‘where/how’) recognizes motion and locates objects



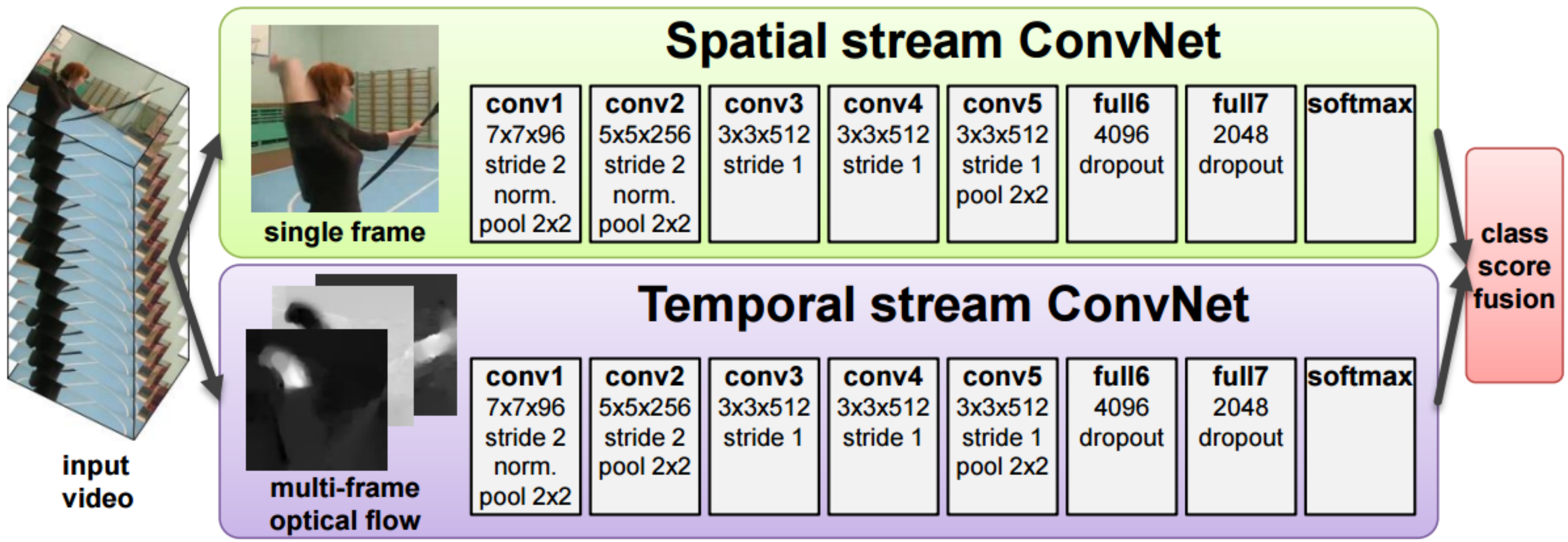


➔ **“Interconnection”**
e.g. in STS area

➔ Ventral (‘what’) stream performs object recognition

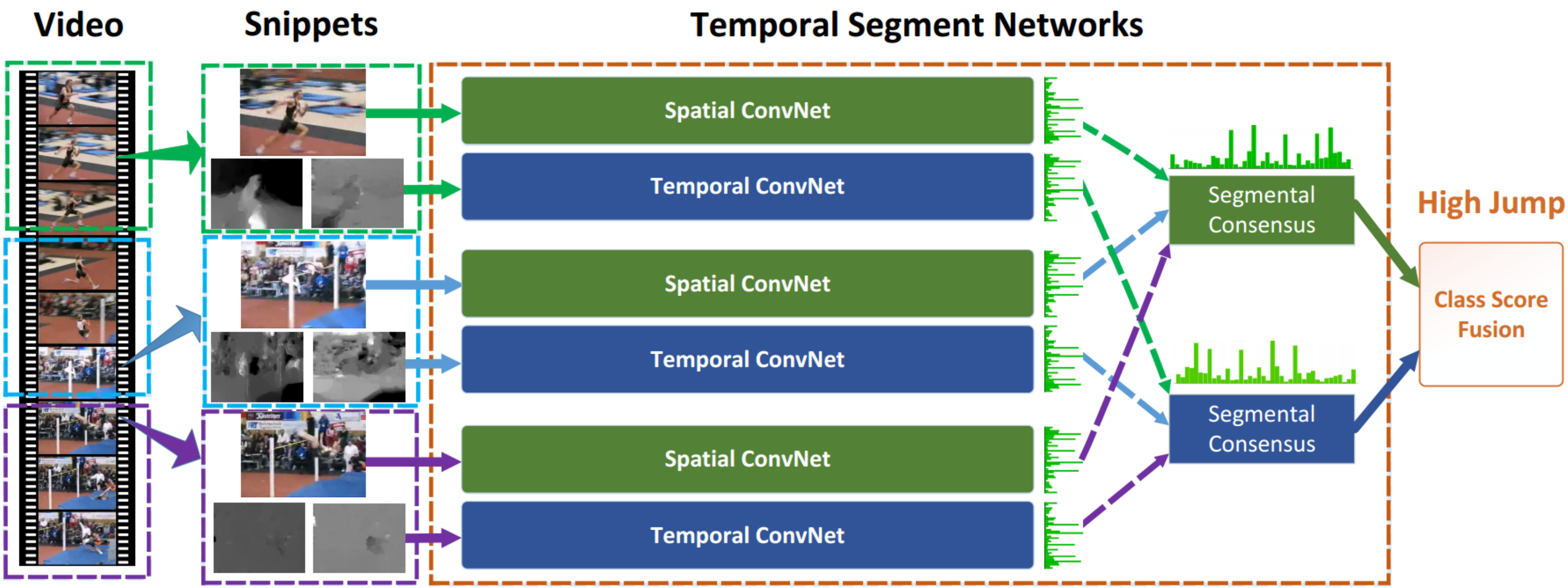
Sources: *“Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli." Journal of neurophysiology 65.6 (1991).*
“A cortical representation of the local visual environment”, Nature. 392 (6676): 598–601, 2009
https://en.wikipedia.org/wiki/Two-streams_hypothesis

2-Stream Network



Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

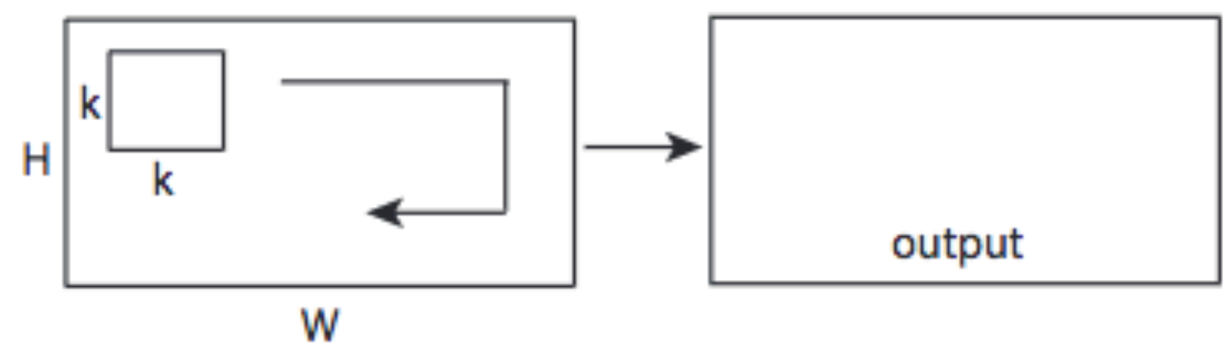
Temporal segment network



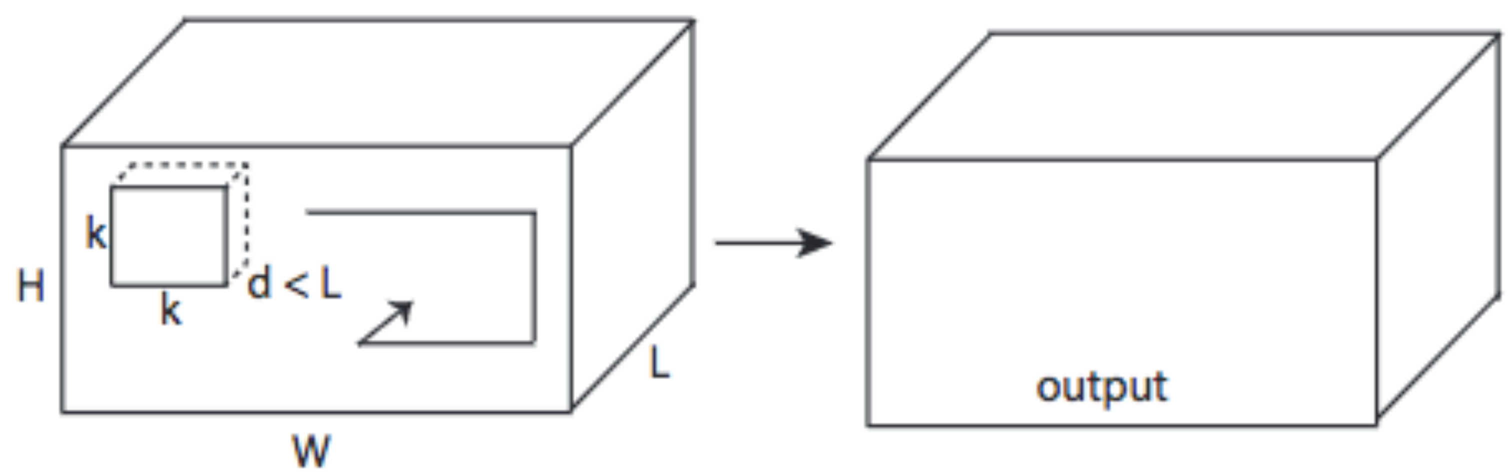
Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, ECCV 2016

3D convolutional Networks

2D convolutions



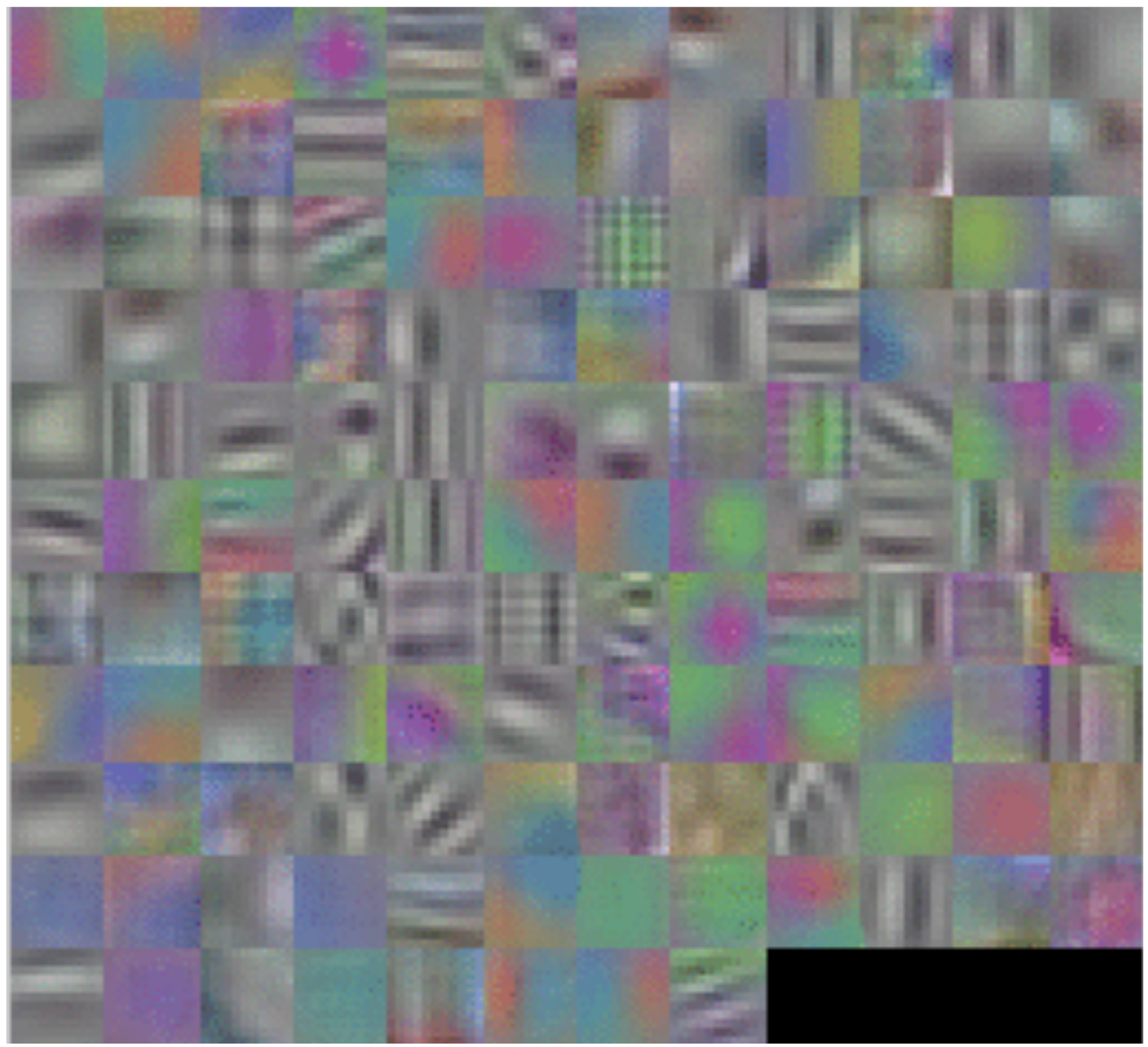
3D convolutions



Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

3D convolutional Networks

- 3D filters at the first layer.

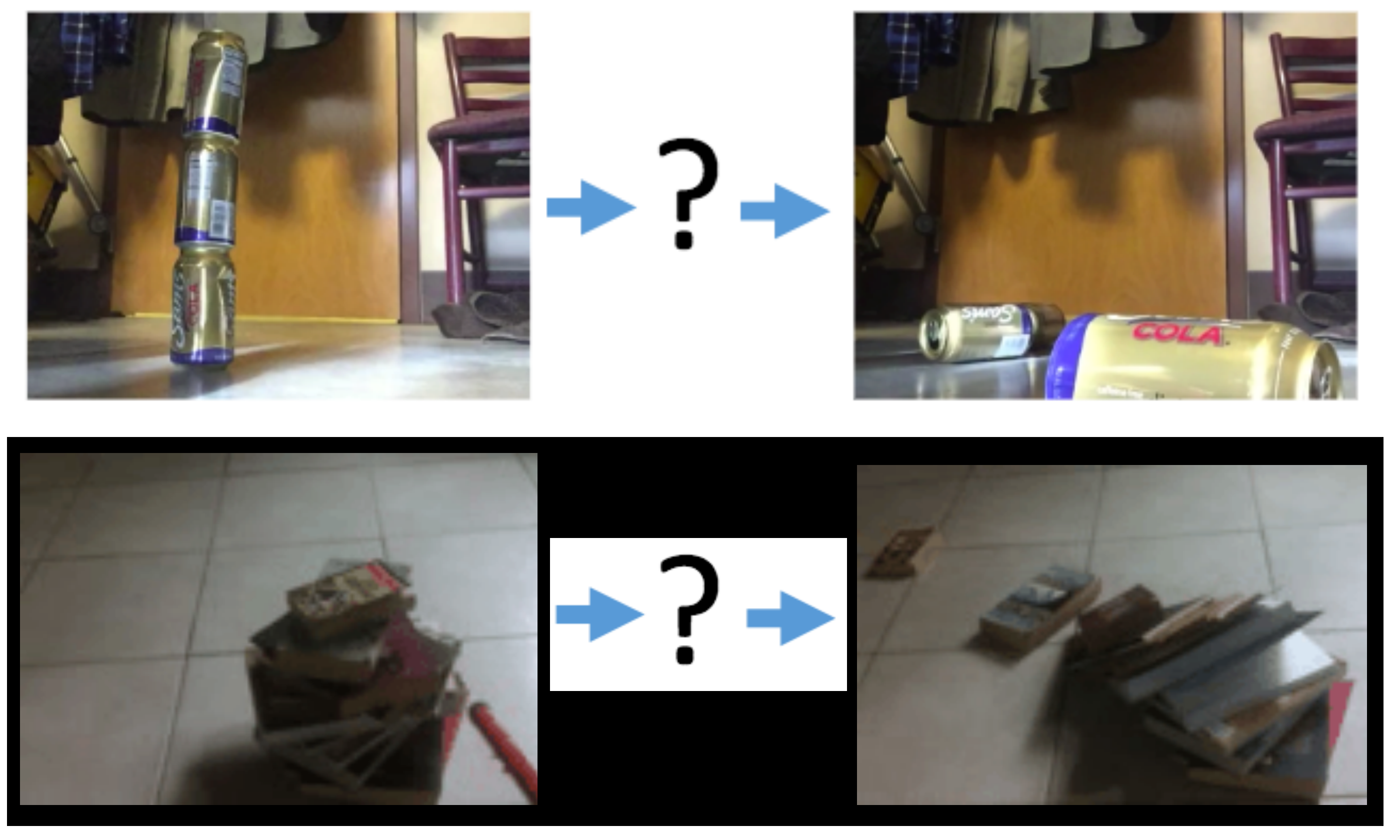


Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

Temporal Relational Reasoning

- Infer the temporal relation between frames.

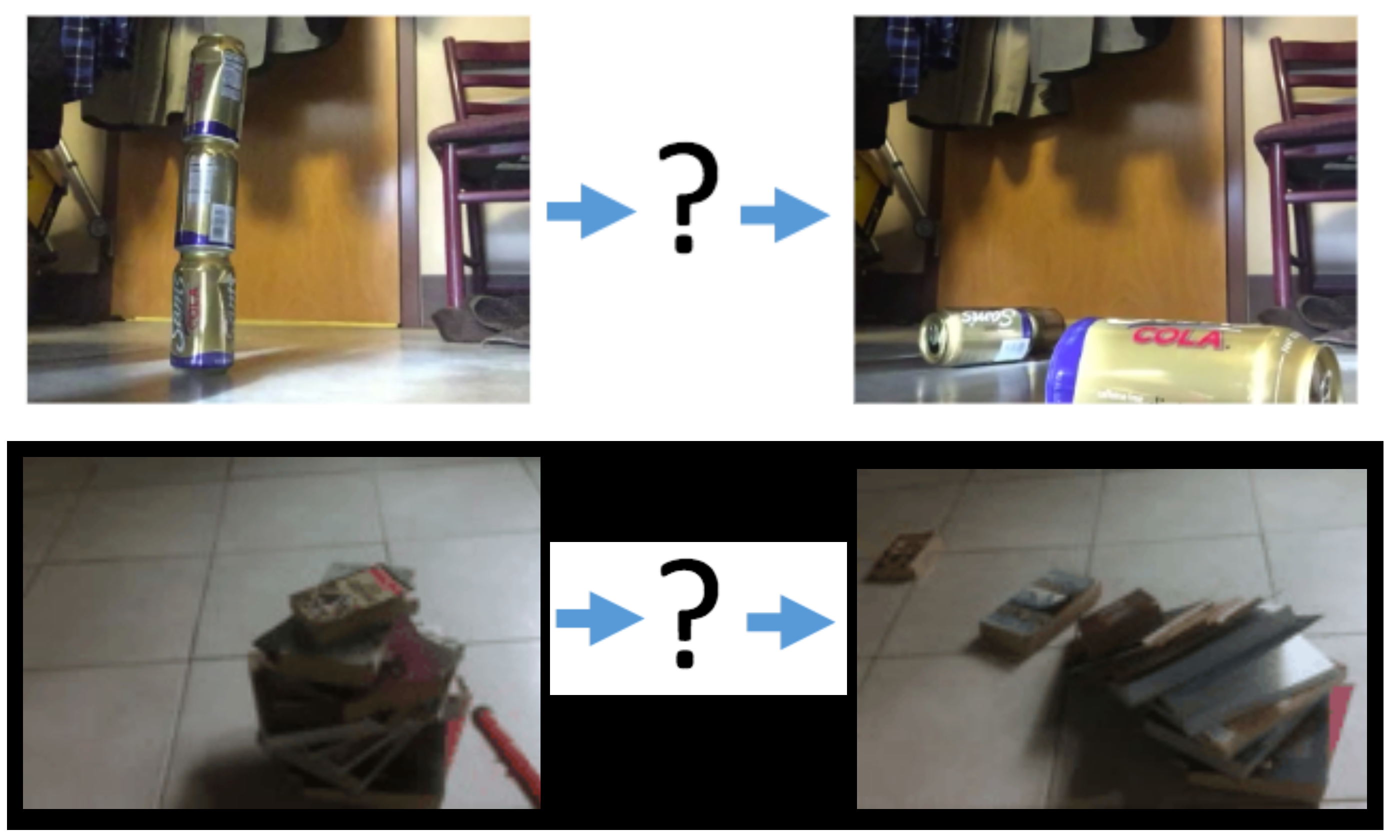
Poking a stack of something so it collapses



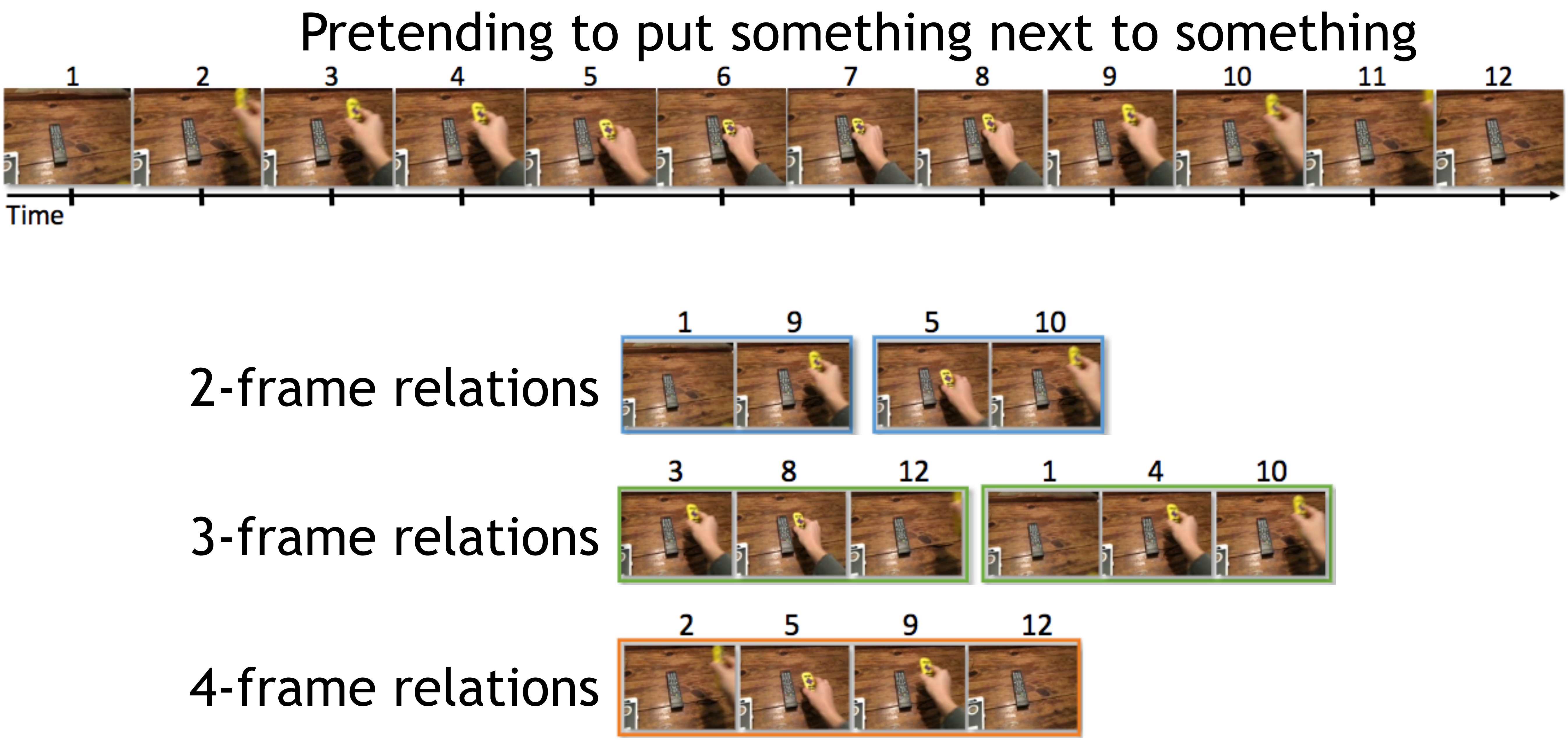
Temporal Relational Reasoning

- It is the **temporal transformation/relation** that defines the activity, rather than the **appearance of objects**.

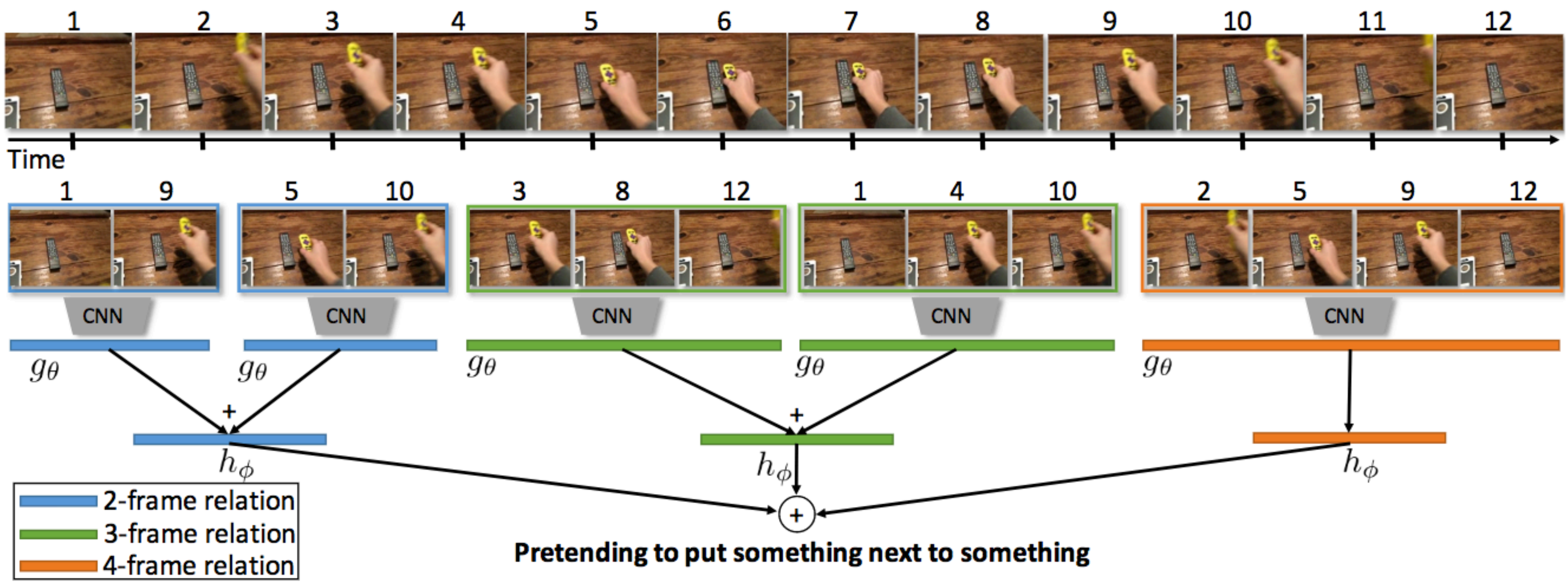
Poking a stack of something so it collapses



Temporal Relations in Videos



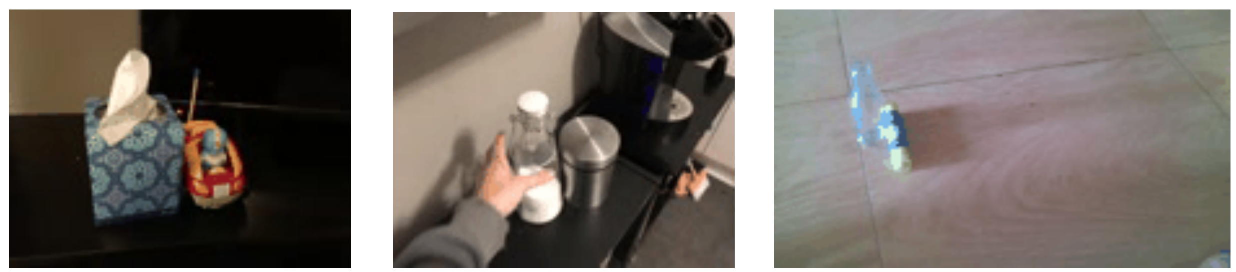
Framework of Temporal Relation Networks



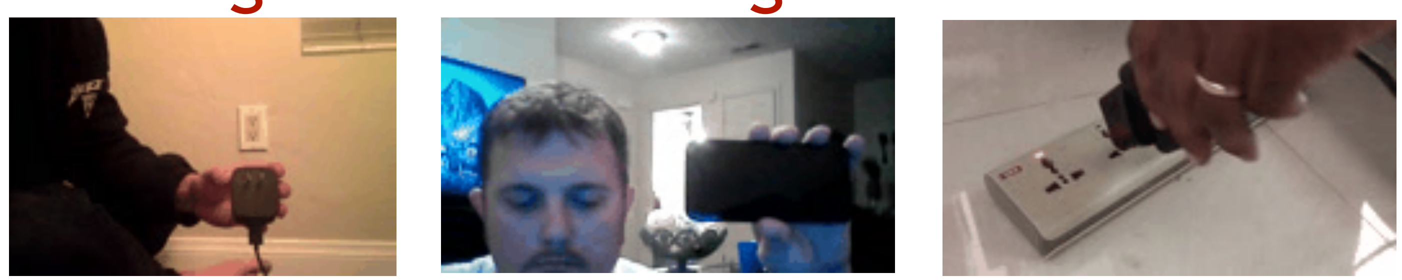
Something-Something Dataset

- 100 K videos from 174 human-object interaction classes.

Moving something away from something



Plugging something into something

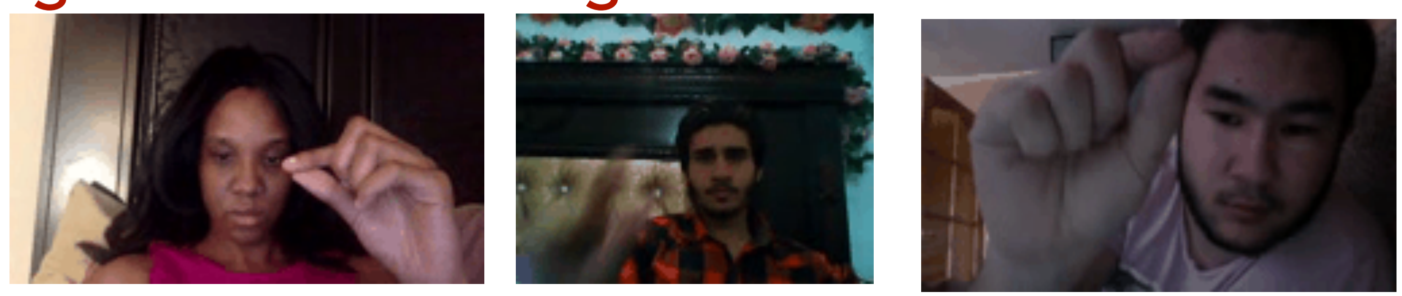


Pulling two ends of something so that it gets stretched

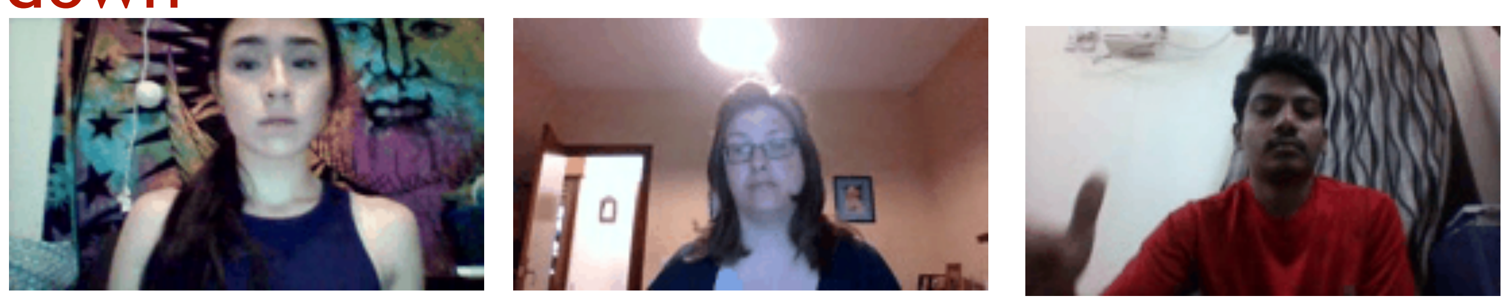
Jester Dataset

- 140 K videos from 27 gesture classes.

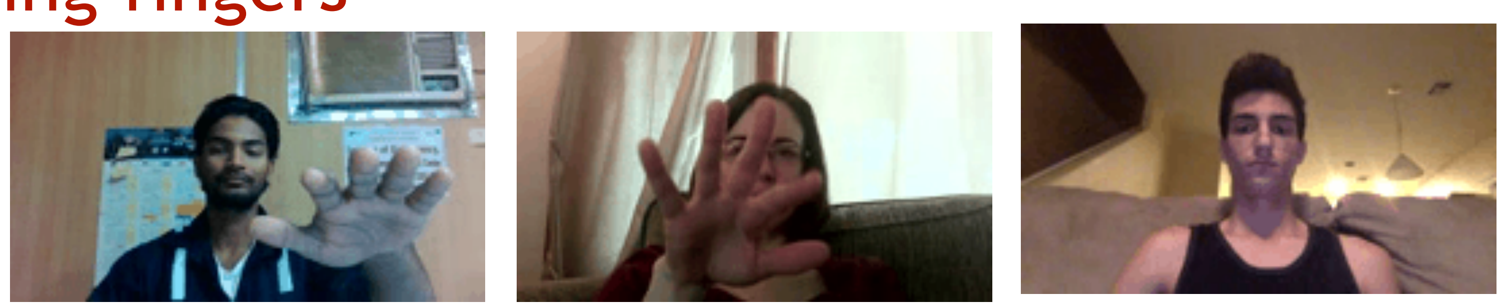
Zooming in with two fingers



Thumb down



Drumming fingers



Experimental Results

- On Something-Something dataset

model	Top1 acc.(%)	Top5 acc.
single frame	11.41	33.39
2-frame TRN	22.23	48.80
3-frame TRN	26.22	54.15
4-frame TRN	29.83	58.21
5-frame TRN	30.39	58.29
7-frame TRN	31.01	59.24
MultiScale TRN	33.01	61.27
MultiScale TRN (10-crop)	**34.44**	**63.20**

model	Top1 acc.(%)
Yana Hasson	25.55
Harrison.AI	26.38
I3D by [8]	27.23
Guillaume Berger	30.48
Besnet (Top1 on leaderboard)	31.66
MultiScale TRN	**33.60**

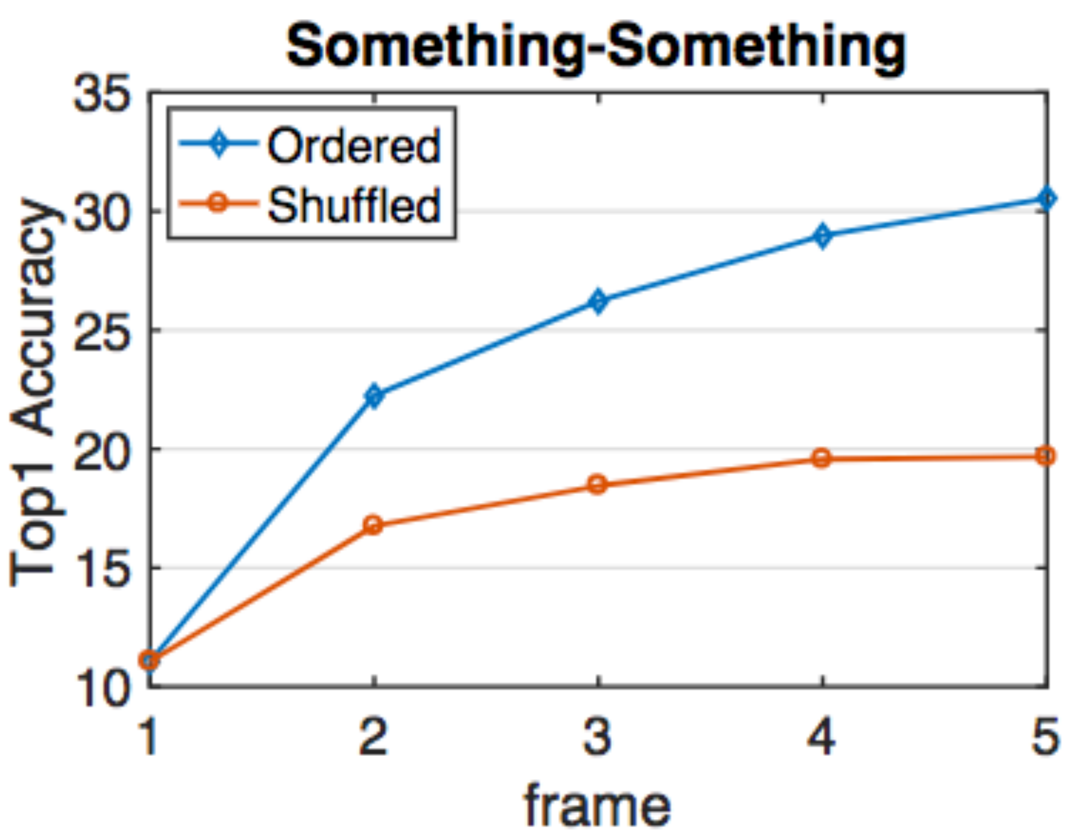
Experimental Results

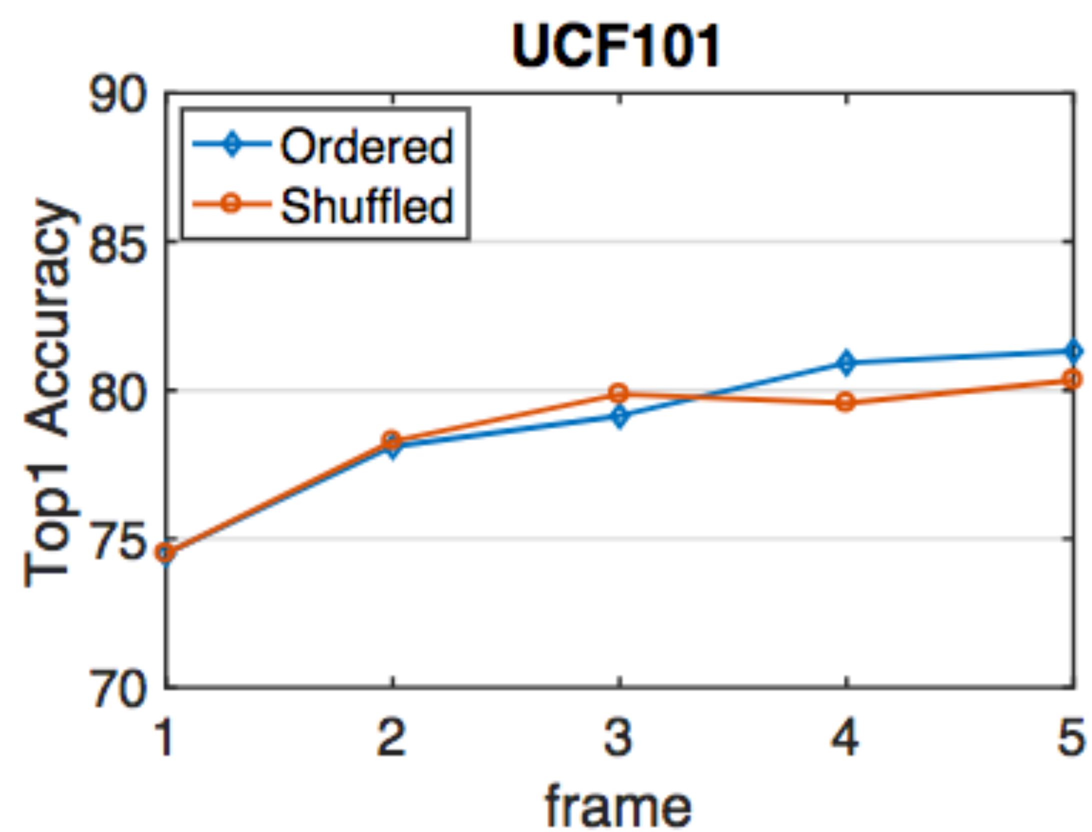
- On Jester dataset

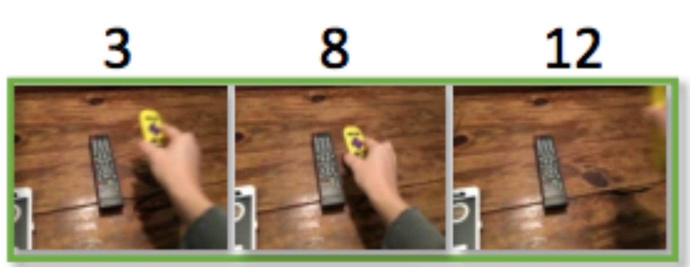
model	Top1 acc.(%)	Top5 acc.
single frame	63.60	92.44
2-frame TRN	75.65	94.40
MultiScale TRN	93.70	99.59
MultiScale TRN (10-crop)	**95.31**	**99.86**

model	Top1 acc.(%)
20BN's Jester System	82.34
VideoLSTM	85.86
Guillaume Berger	93.87
Ford's Gesture Recognition System	94.11
Besnet (Top1 on leaderboard)	94.23
MultiScale TRN	**94.78**

Importance of temporal orders







How well are they diving?

Olympic judge's score

Pirsiavash, Vondrick, Torralba. Assessing Quality of Actions, ECCV 2014

<u>How well</u> **are they diving?**

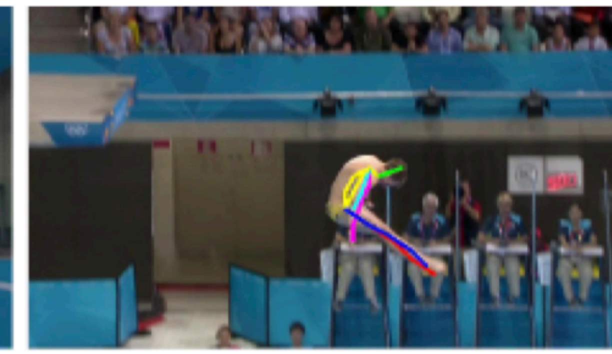
1. Track and compute human pose

<u>How well</u> **are they diving?**

1. Track and compute human pose

2. Extract temporal features
   - take FT and histogram?
   - use deep network?

<u>How well</u> are they diving?

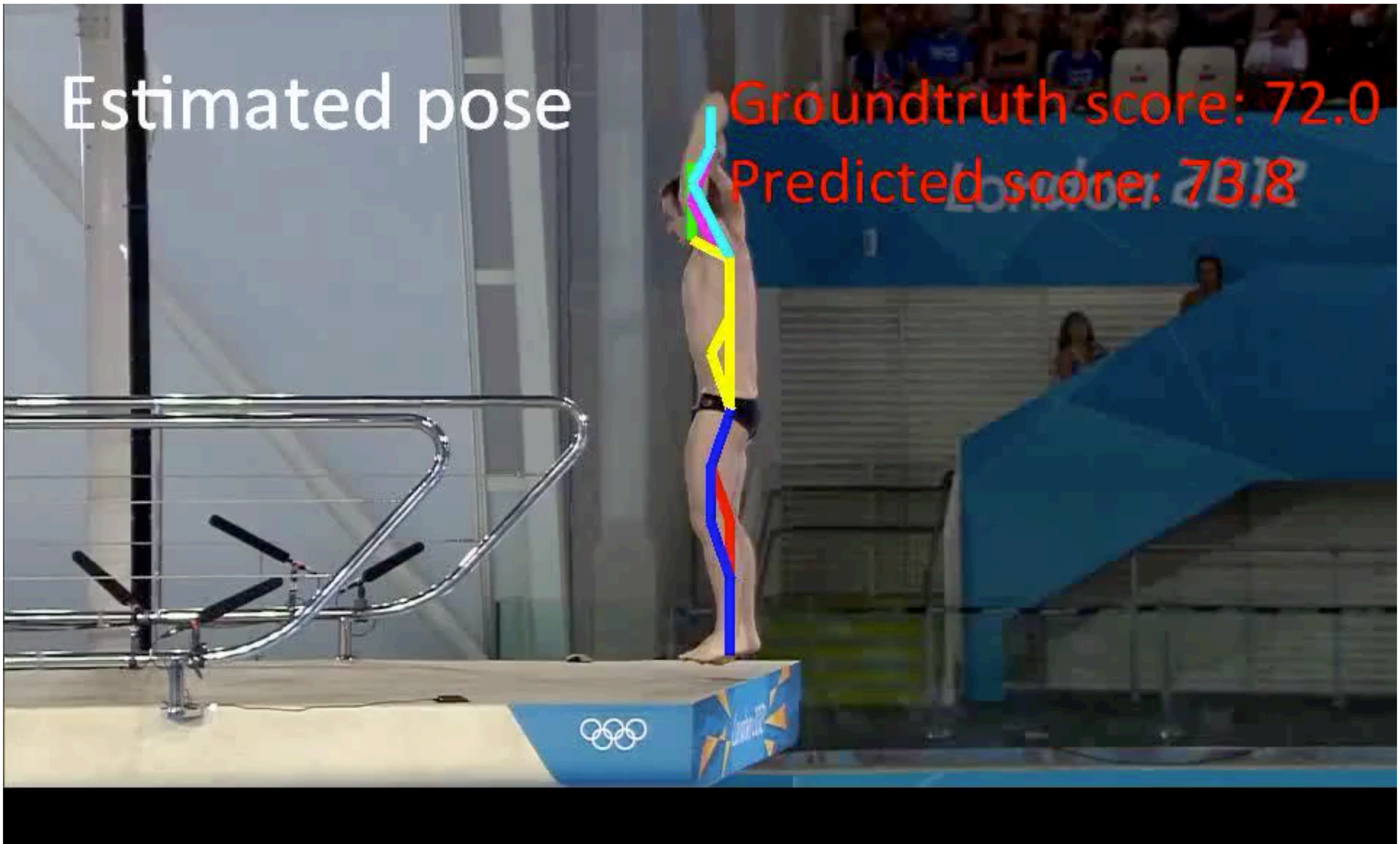


1. Track and compute human pose

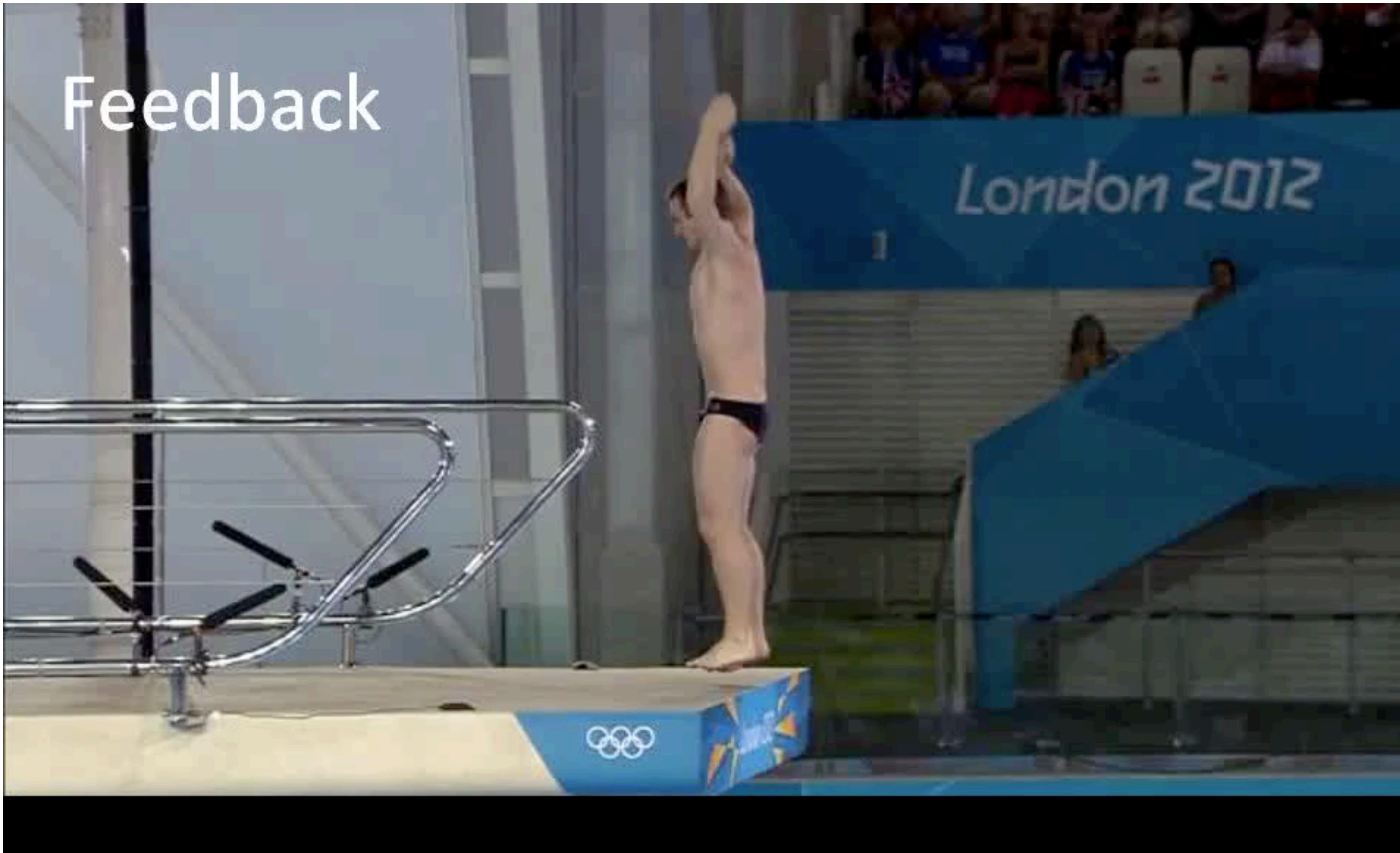
2. Extract temporal features
   - take FT and histogram?
   - use deep network?

3. Train regression model to predict expert quality score

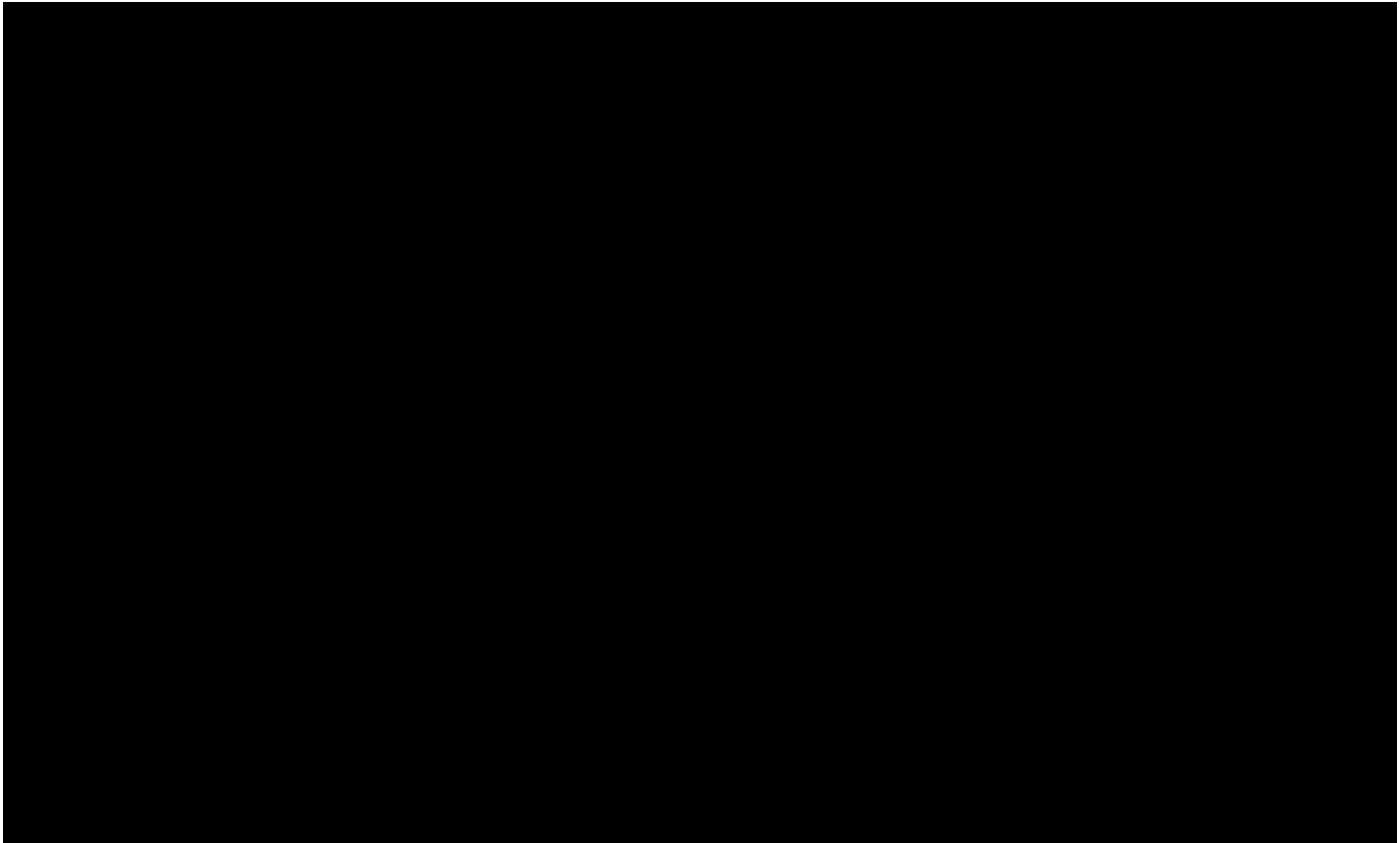
Assessing diving



Feedback



Summarizing



Assessing figure skating