Video Classification

Danna Gurari
University of Colorado Boulder
Fall 2021

Review

- Last week's lecture topic:
  - Salient object detection

- Assignments (Canvas)
  - Reading assignment due earlier today
  - Reading assignments due Wednesday and next week
  - Final project proposal due in just over 3 weeks

- Questions?

Final Project Requirements

- Described on the course website
  - https://home.cs.colorado.edu/~DrG/Courses/RecentAdvancesInComputerVision/FinalProject.html
- Multiple milestones
  - Project proposal
  - Project outline
  - Final project presentation
  - Peer evaluation
  - Final report

Video Classification: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models

Video Classification: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models

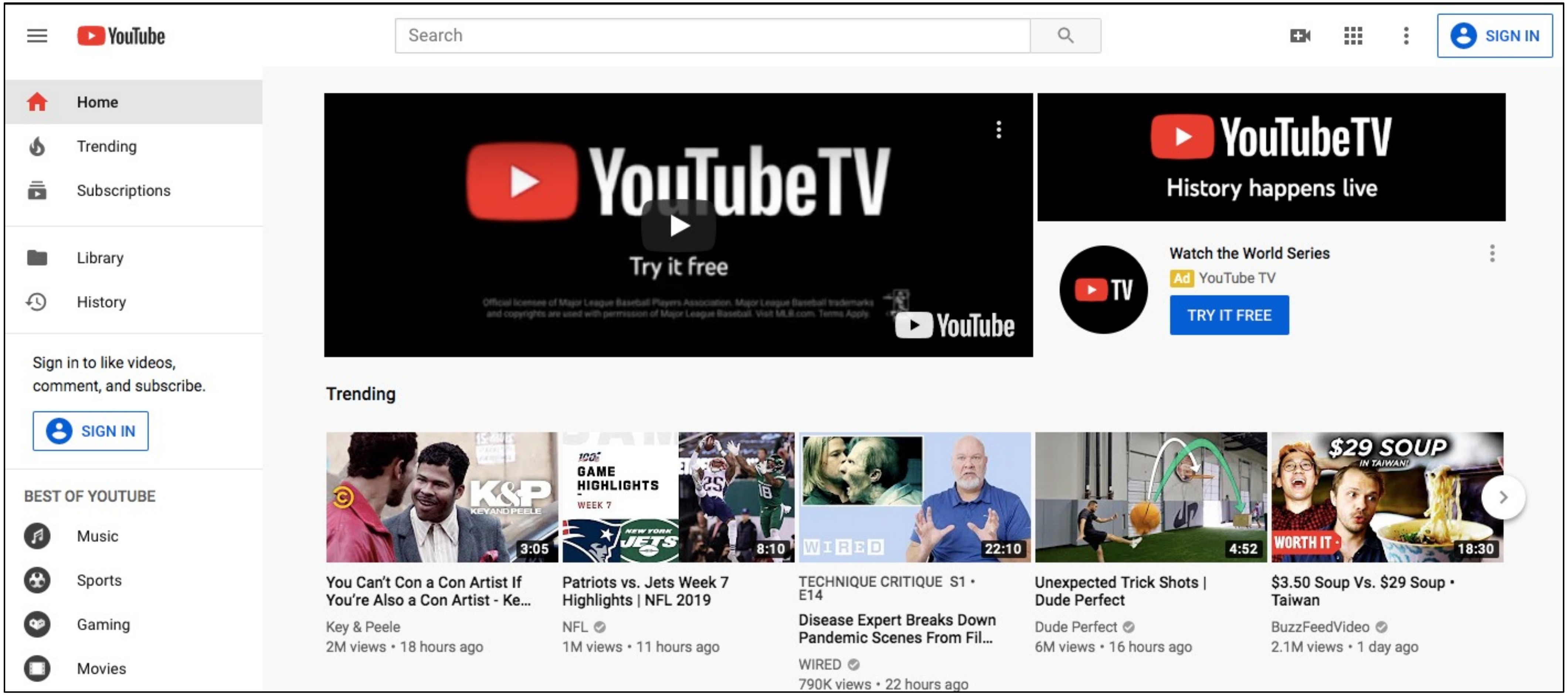
Definition

- Assign a video a label from a set of categories; typically, multiple choice but also can be multiple labels
  - e.g., activity or topical themes

Video Classification/Localization: Today’s Topics

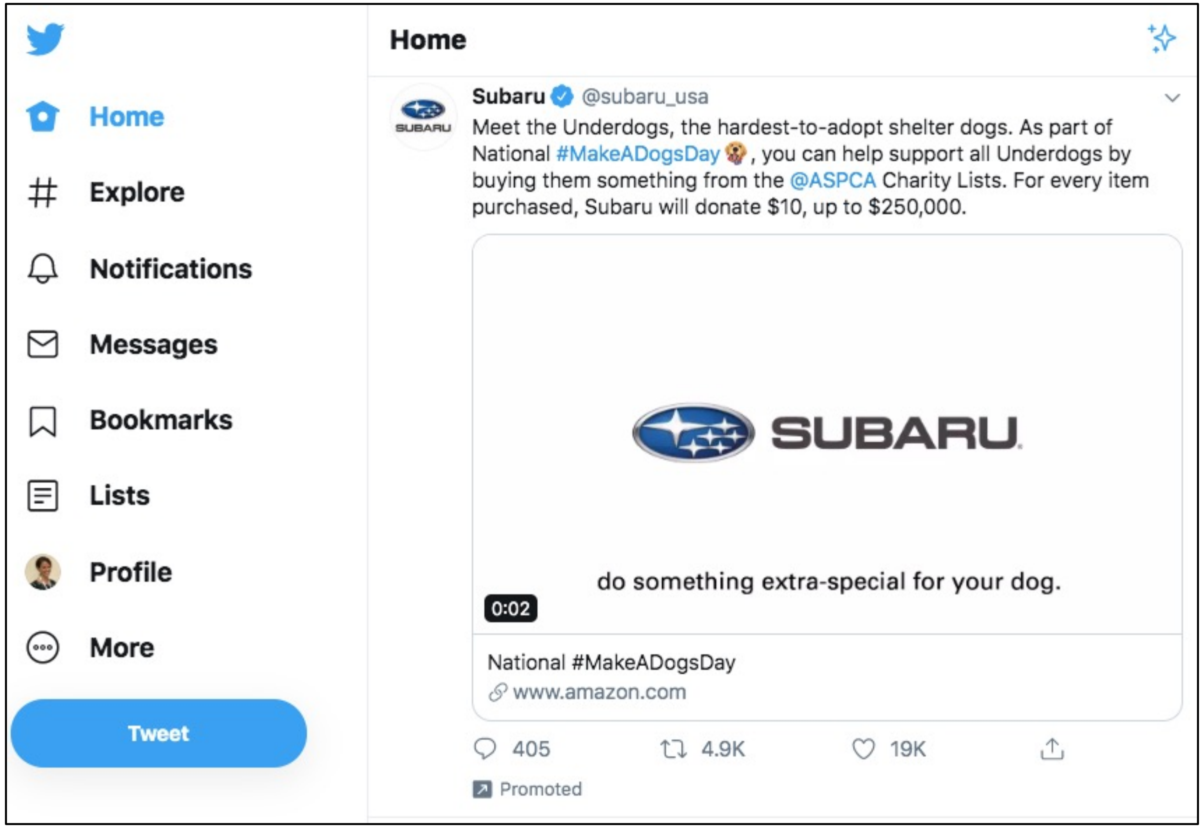
- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models

Video Search



300 hours of video uploaded every minute (https://merchdope.com/youtube-stats/)

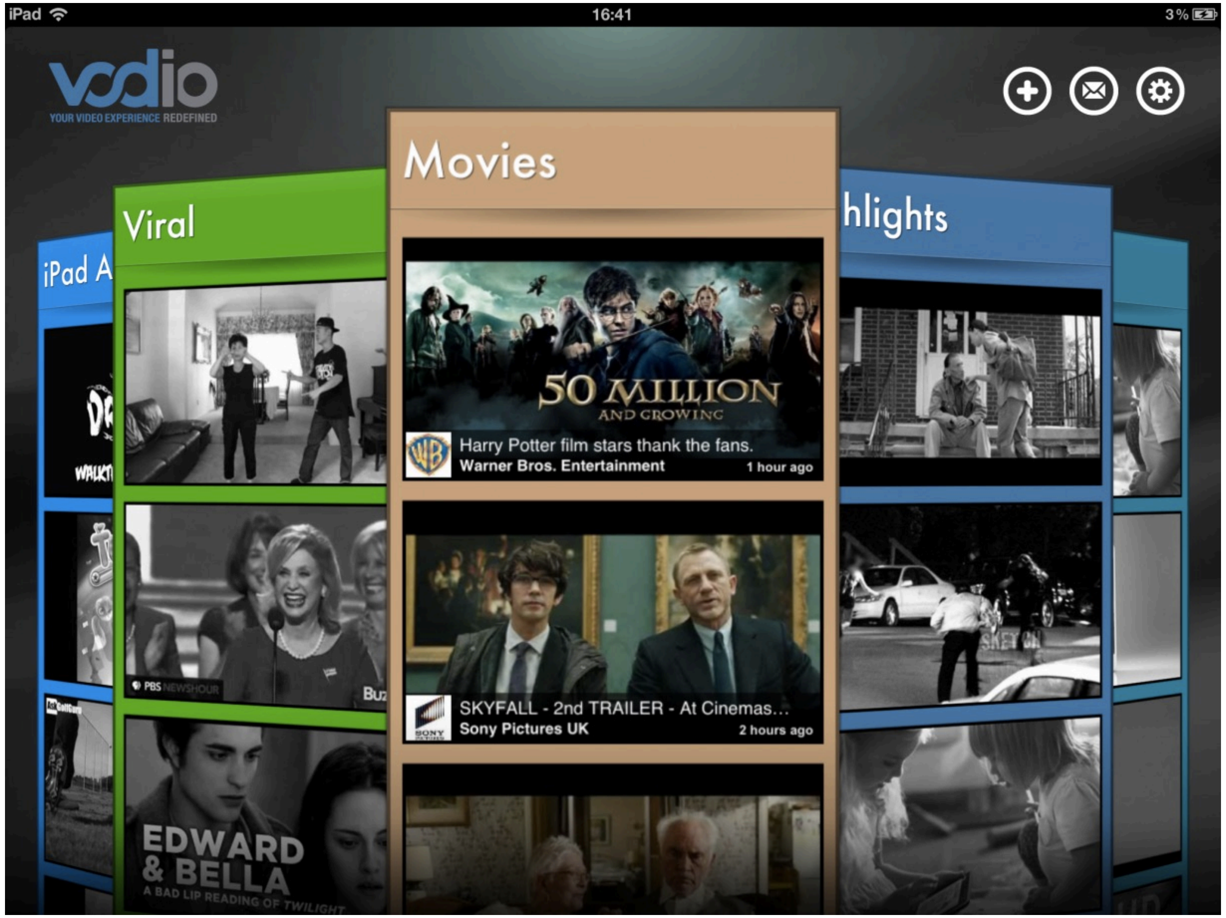
Social Media Recommendations



“An estimated 12 million micro-videos are posted to Twitter each day. The number of microvideos produced surpasses the total inventory of YouTube every 3 months”

- “The Open World of Micro-Videos; Nguyen et al.; https://www.ics.uci.edu/~fowlkes/papers/nrfr_bigvision.pdf

Video Organization



Lists search results based on your collection of videos (spanning YouTube, news, movies, and more) in one list

Automatically Remove Objectionable Content

Nudity or sexual content

YouTube is not for pornography or sexually explicit content. If this describes your video, even if it's a video of yourself, don't post it on YouTube. Also, be advised that we work closely with law enforcement and we report child exploitation. Learn more

Harmful or dangerous content

Don't post videos that encourage others to do things that might cause them to get badly hurt, especially kids. Videos showing such harmful or dangerous acts may get age-restricted or removed depending on their severity. Learn more

Hateful content

Our products are platforms for free expression. But we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line. Learn more

Violent or graphic content

It's not okay to post violent or gory content that's primarily intended to be shocking, sensational, or gratuitous. If posting graphic content in a news or documentary context, please be mindful to provide enough information to help people understand what's going on in the video. Don't encourage others to commit specific acts of violence. Learn more

And more listed here: https://www.youtube.com/about/policies/#community-guidelines

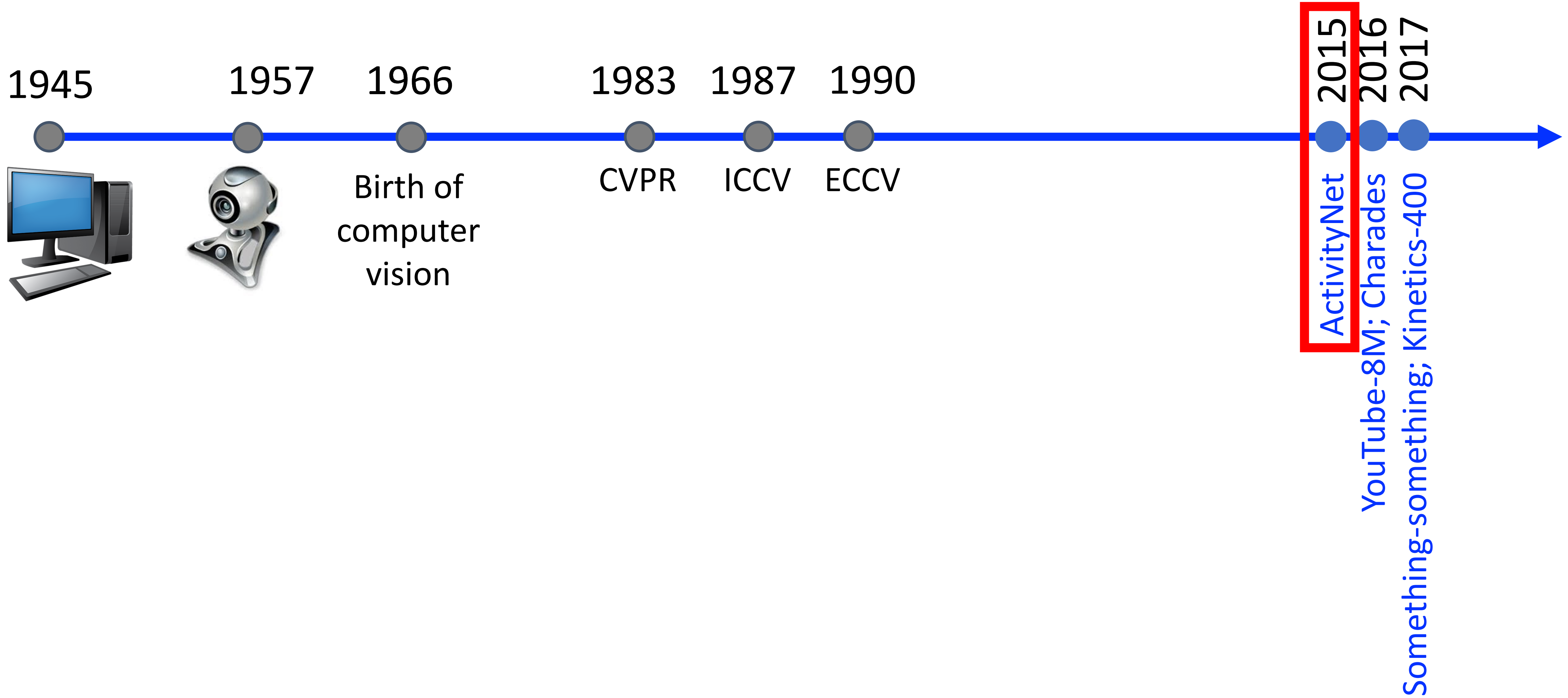
Applications

For what other applications might video classification be useful?

Video Classification: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models

Datasets



1945
1957
1966
Birth of computer vision
1983
CVPR
1987
ICCV
1990
ECCV
2015
ActivityNet
2016
YouTube-8M; Charades
2017
Something-something; Kinetics-400

ActivityNet

Focus on activities that humans spend most of their time doing in their lives

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. CVPR 2015.

ActivityNet

1. Category Selection

* American Time Use Survey (ATUS) created by the Department of Labor organizes activities according to:
- social interactions
- where activity usually occurs

* Authors selected 203 from the 2000+ activities in ATUS:
- 7 top-level categories: *Personal Care*, *Eating and Drinking*, *Household*, *Working*,…
- 4-level hierarchy

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. CVPR 2015.

ActivityNet

1. Category Selection

* American Time Use Survey (ATUS) created by the Department of Labor organizes activities according to:
- social interactions
- where activity usually occurs

* Authors selected 203 from the 2000+ activities in ATUS:
- 7 top-level categories: *Personal Care*, *Eating and Drinking*, *Household*, *Working*,…
- 4-level hierarchy

2. Video Collection

YouTube

3. Video Verification

* “Expert” AMT workers verify presence of activity in each video

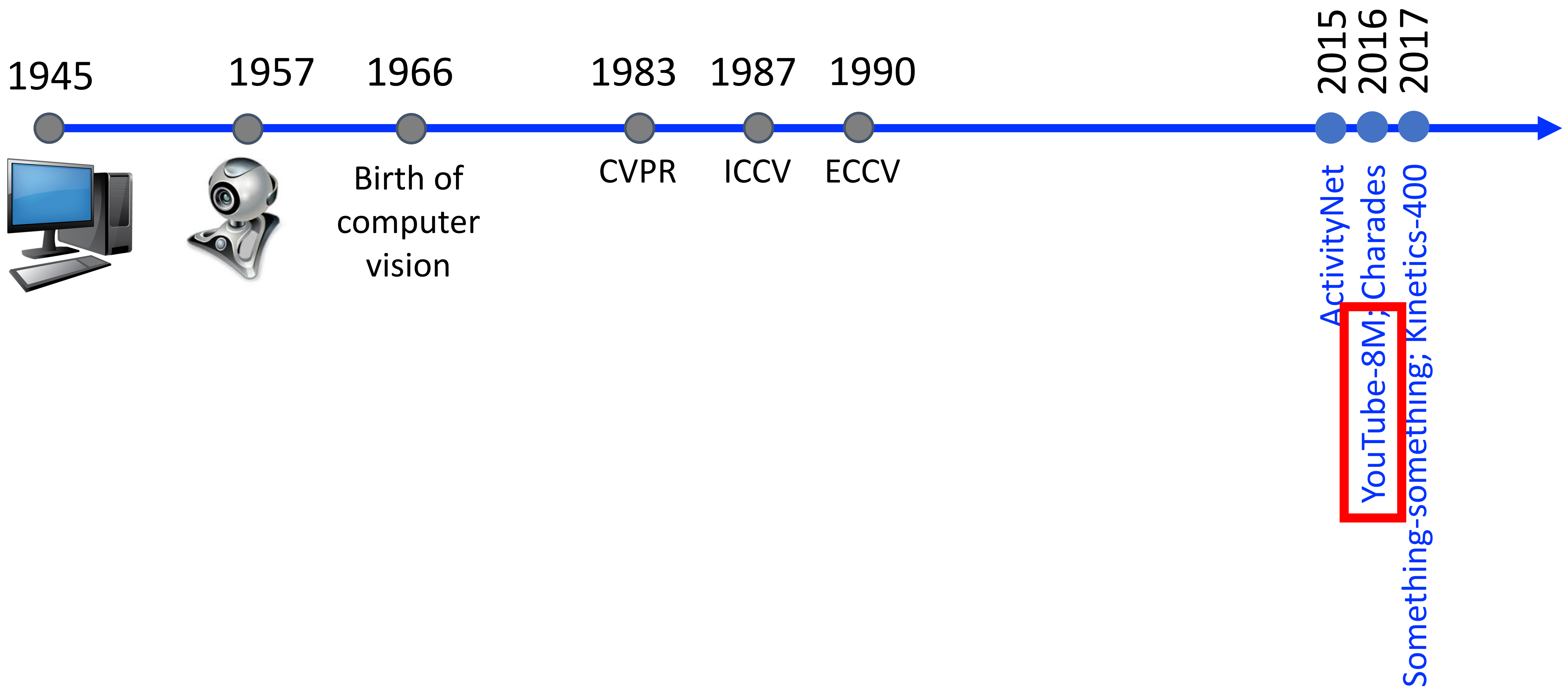
* Honey pot tasks introduced to assess trust of each crowd worker’s work

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. CVPR 2015.

ActivityNet Workshop

http://activity-net.org/challenges/2021/

Datasets



1945
1957
1966
Birth of computer vision
1983
CVPR
1987
ICCV
1990
ECCV
2015
ActivityNet
2016
YouTube-8M; Charades
2017
Something-something; Kinetics-400

YouTube-8M

Largest multi-label video classification dataset for determining the key topical themes of the video

- $\sim$8 million videos of over 500,000 hours
- 4,800 classes spanning "activities (sports, games, hobbies), objects (autos, food, products), scenes (travel), and events"

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. arXiv 2016.

YouTube-8M

1. Category Selection

* Starting point: 50,000 video topics from a knowledge graph (Freebase); e.g., people, places

* Kept ~10,000 topics that most of 3 humans indicated are visually distinguishable and do not require domain expertise to recognize

* Reduced to categories that are popular: 1,000+ views, > 120 secs, < 500 secs, and >= 200 videos

Manual pruning task:

Entity Name	Entity URL	Entity Description
Thunderstorm	http://www.freebase.com/m/0jb2l	A thunderstorm, also known as an electrical storm, a lightning storm, or a thundershower, is a type of storm characterized by the presence of lightning and its acoustic effect on the Earth's atmosphere known as thunder. The meteorologically assigned cloud type associated with the thunderstorm is the cumulonimbus. Thunderstorms are usually accompanied by strong winds, heavy rain and sometimes snow, sleet, hail, or no precipitation at all...

How difficult is it to identify this entity in images or videos (without audio, titles, comments, etc)?

1. Any layperson could
2. Any layperson after studying examples, wikipedia, etc could
3. Experts in some field can
4. Not possible without non-visual knowledge
5. Non-visual

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. arXiv 2016.

YouTube-8M

1. Category Selection

* Starting point: 50,000 video topics from a knowledge graph (Freebase); e.g., people, places

* Kept ~10,000 topics that most of 3 humans indicated are visually distinguishable and do not require domain expertise to recognize

* Reduced to categories that are popular: 1,000+ views, > 120 secs, < 500 secs, and >= 200 videos

2. Video Collection

3. Label Verification

* 3 humans rate labels for 8000 images
* With respect to the human raters
- 78.8% precision
- 14.5% recall
* Inter-rater agreement: ~80%

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, Sudheendra Vijayanarasimhan; arXiv 2016; *10 citations in 3/17*

YouTube-8M Challenge & Annual Workshop

YouTube | 8M   Dataset   Explore   Download   Workshop   About

2019
2018
2017

Updated Dataset

YouTube-8M Segments was released in June 2019 with segment-level annotations. Human-verified labels on about 237K seg
1000 classes are collected from the validation set of the YouTube-8M dataset. Each video will again come with time-localized
features so classifier predictions can be made at segment-level granularity.

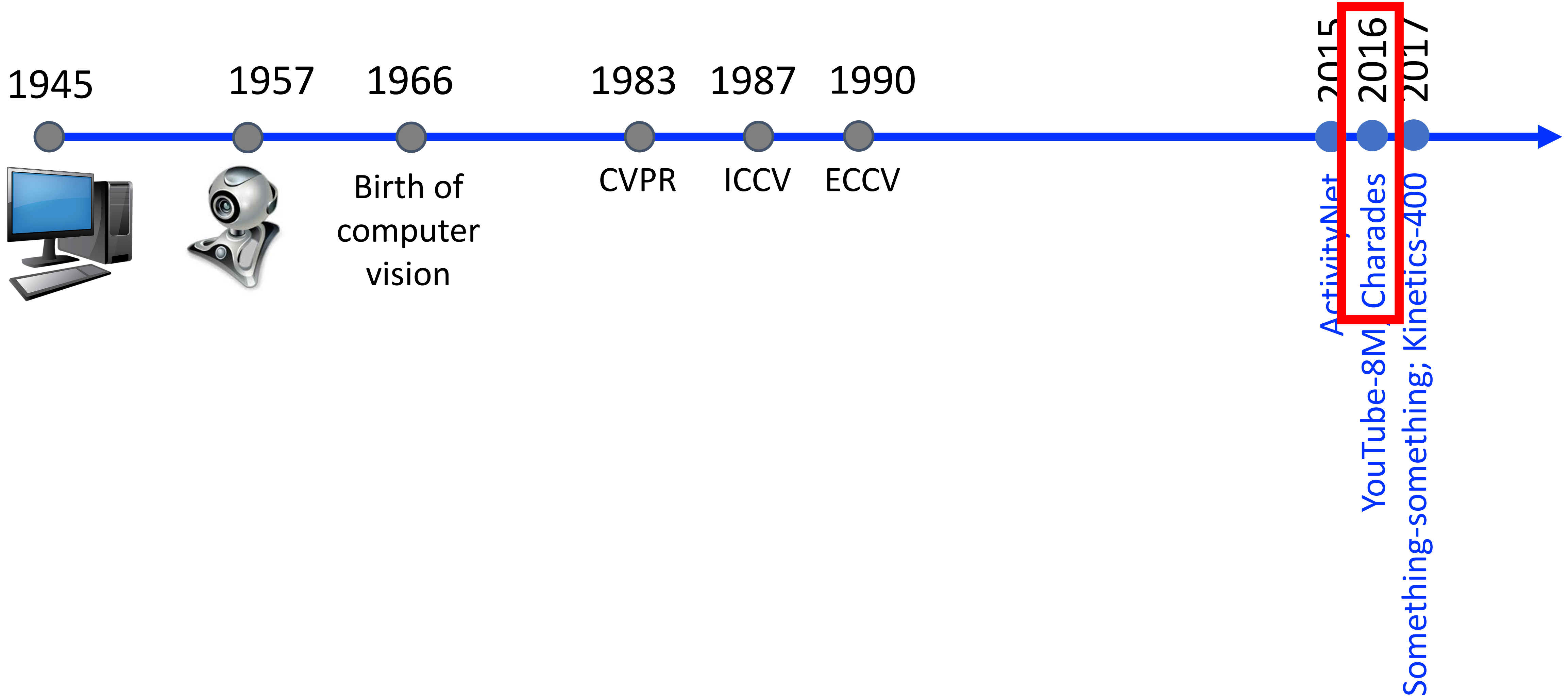
YouTube-8M was updated in May 2018 to include higher-quality, more topical annotations, and to clean up the annotation vocabulary. A number of low-frequency or low-quality labels and associated videos were removed, resulting in a smaller but higher-quality dataset (5.6M videos, 3862 classes). Additionally, the video IDs in the TensorFlow Record files have been anonymized, and the mapping to the real YouTube IDs will be periodically updated to exclude any videos that have been subsequently deleted (while preserving their anonymized features).

Dataset versions:

1. Jun 2019 version (current): 230K human-verified segment labels, 1000 classes, 5 segments/video
2. May 2018 version (current): 6.1M videos, 3862 classes, 3.0 labels/video, 2.6B audio-visual features
3. Feb 2017 version (deprecated): 7.0M videos, 4716 classes, 3.4 labels/video, 3.2B audio-visual features
4. Sep 2016 version (deprecated): 8.2M videos, 4800 classes, 1.8 labels/video, 1.9B visual-only features

https://research.google.com/youtube8m/workshop2019/

Datasets



1945
1957
1966
Birth of computer vision
1983
CVPR
1987
ICCV
1990
ECCV
2015
ActivityNet
2016
YouTube-8M; Charades
2017
Something-something; Kinetics-400

Charades

Collection of “boring” videos reflecting daily lives

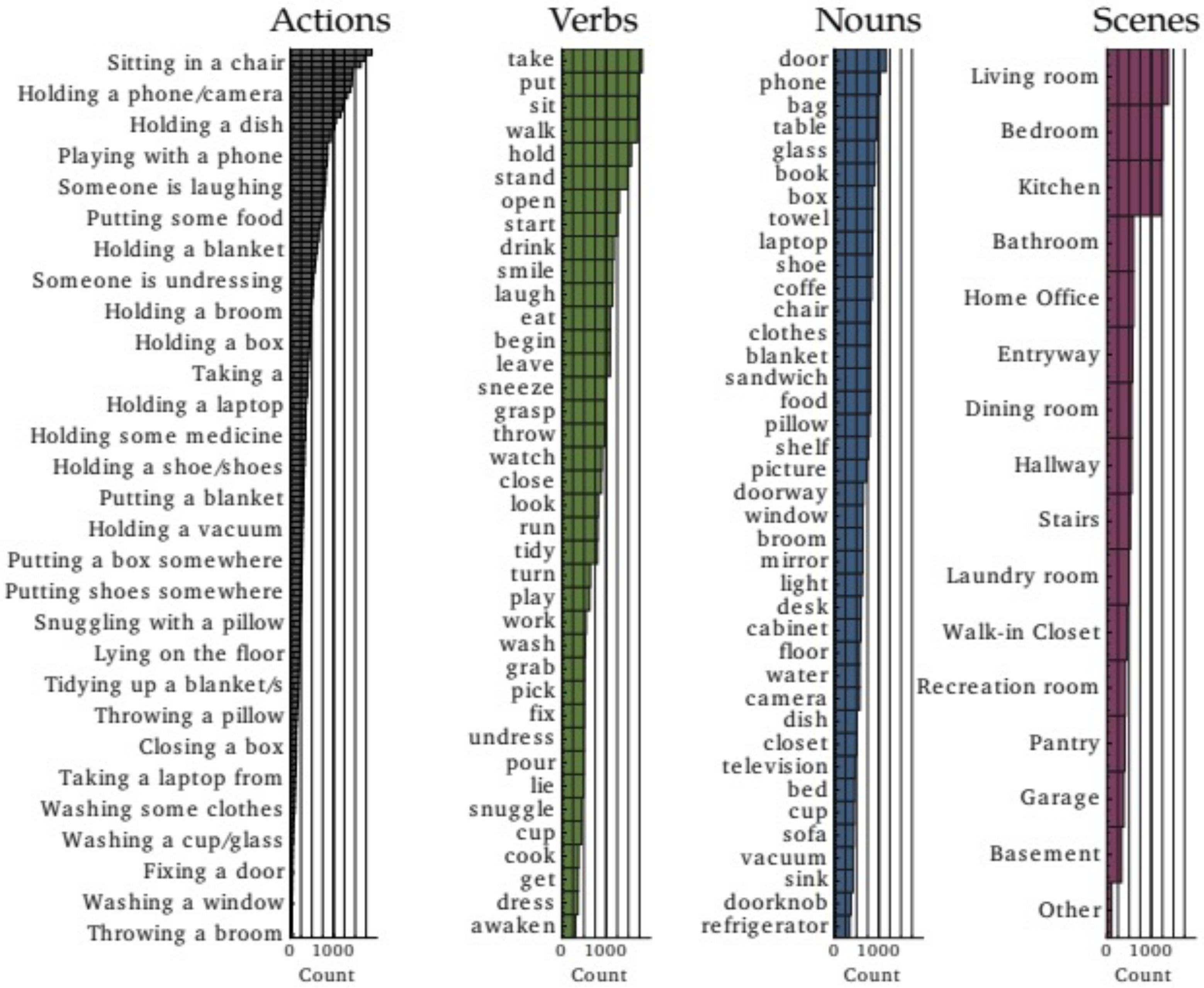
- $\sim$9,848 videos with average length of 30 seconds
- Activities of 267 people from three continents
- Annotations include action labels and classes of interacted objects

Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV 2016.

Charades

1. Video Script Generation

* Authors identified 15 indoor scenes in residential homes (e.g., living room, home office)
* Most common nouns and verbs in these scenes analyzed from 549 movie scripts resulting in 40 objects and 30 actions
* Crowd workers generated scripts describing commonplace, realistic activities that involve 2 objects & 2 actions (given a scene, 5 objects, & 5 actions)



Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV 2016.

Charades

1. Video Script Generation

* Authors identified 15 indoor scenes in residential homes (e.g., living room, home office)
* Most common nouns and verbs in these scenes analyzed from 549 movie scripts resulting in 40 objects and 30 actions
* Crowd workers generated scripts describing commonplace, realistic activities that involve 2 objects & 2 actions (given a scene, 5 objects, & 5 actions)

→

2. Video Collection

* Crowd workers recruited to record 30s videos of them executing the scripts

Demo of videos
https://www.youtube.com/watch?v=x9AhZLDkbyc

Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV 2016.

Charades

1. Video Script Generation

* Authors identified 15 indoor scenes in residential homes (e.g., living room, home office)
* Most common nouns and verbs in these scenes analyzed from 549 movie scripts resulting in 40 objects and 30 actions
* Crowd workers generated scripts describing commonplace, realistic activities that involve 2 objects & 2 actions (given a scene, 5 objects, & 5 actions)

2. Video Collection

* Crowd workers recruited to record 30s videos of them executing the scripts

3. Category Selection

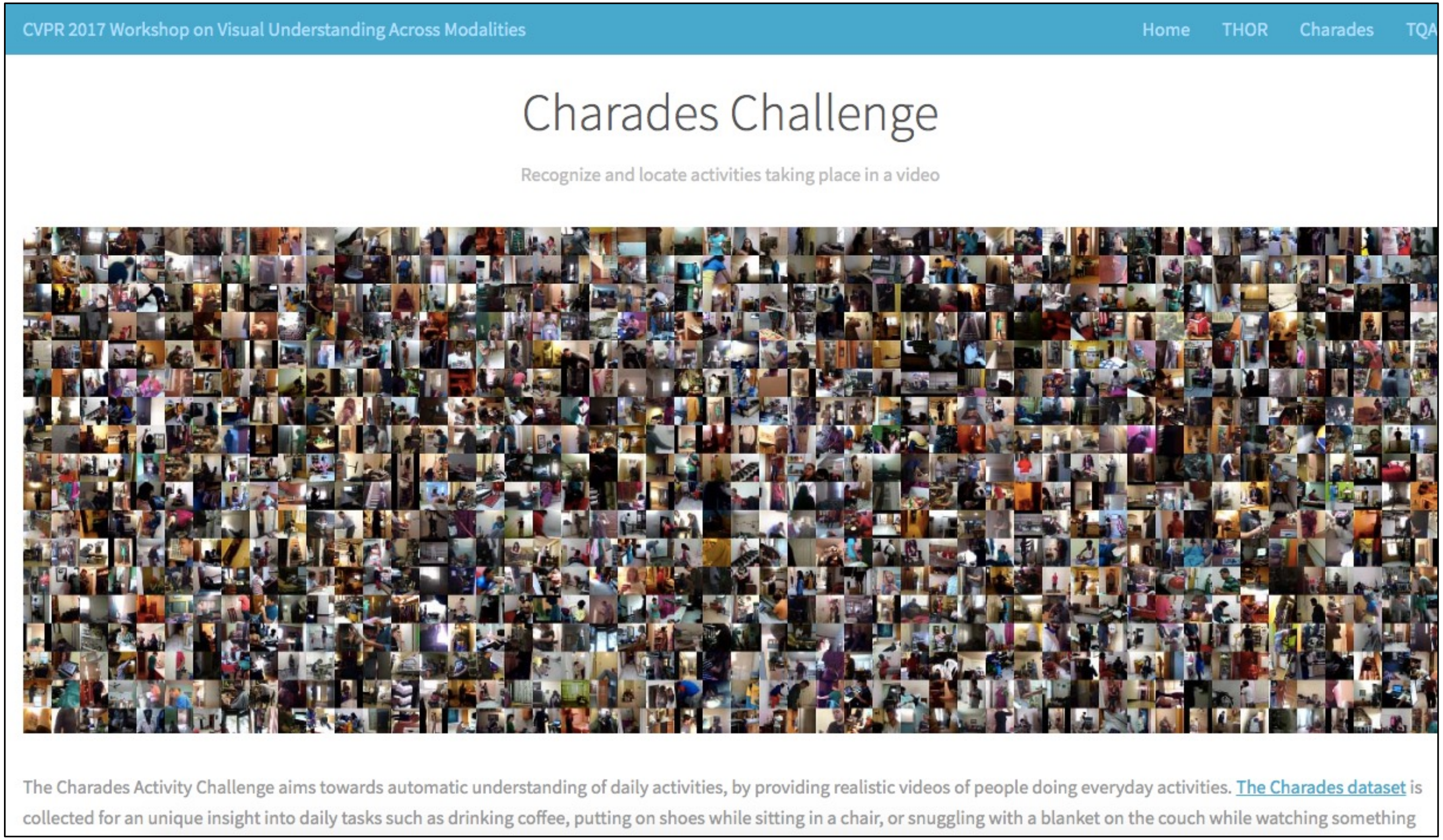
* AMT workers recruited to watch each video and create a description

* Automatically identified “interacted objects” mentioned both in script & description

* 150 actions chosen following crowd worker verification

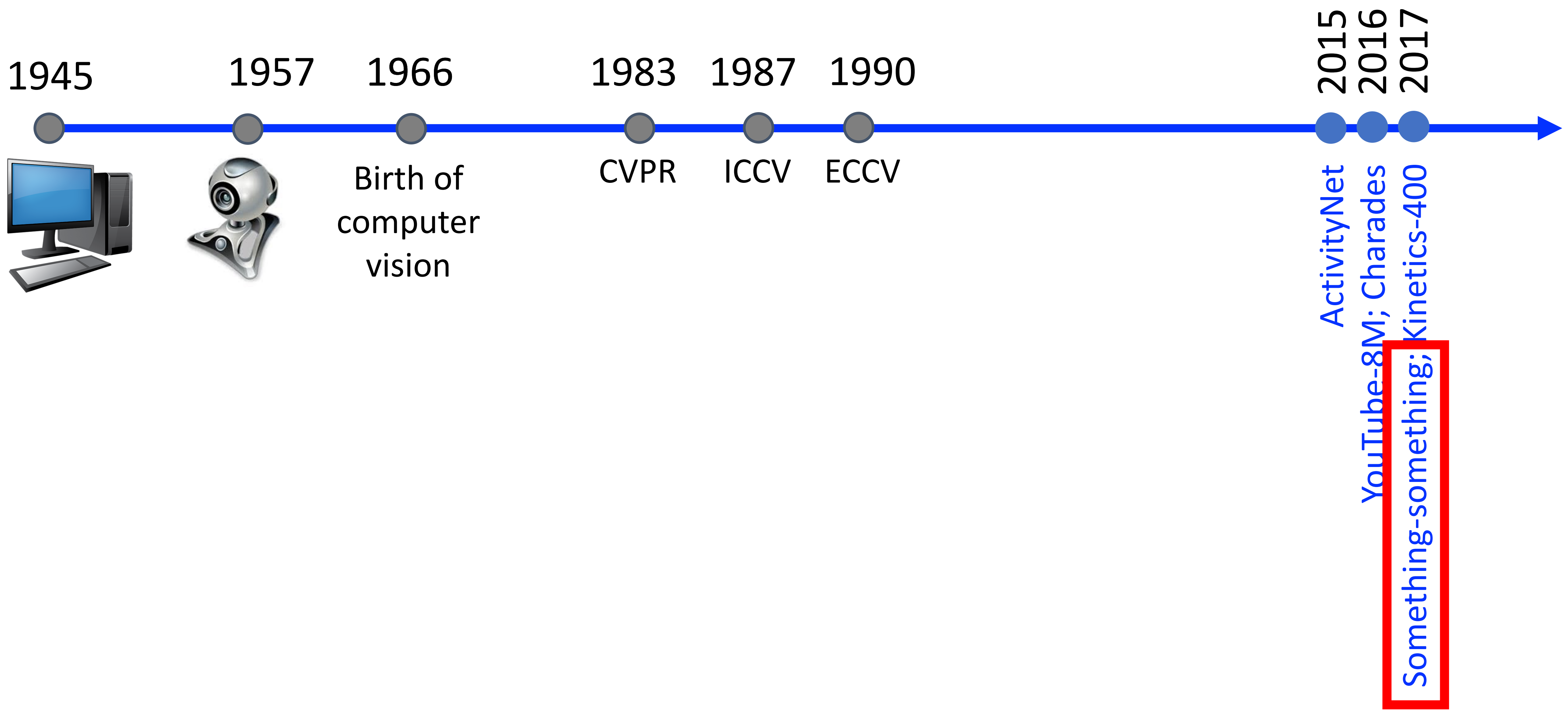
Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV 2016.

Charades Challenge & Annual Workshop



http://vuchallenge.org/charades.html

Datasets



1945
1957
1966
Birth of computer vision
1983
CVPR
1987
ICCV
1990
ECCV
2015
ActivityNet
2016
YouTube-8M; Charades
2017
Something-something; Kinetics-400

Charades

Collection of videos to help models learn common sense features for predicting an activity label; (e.g., “opening” for blinds, door, mouth, zipper)

- More than 100,000 videos ranging from 2 to 6 seconds
- Represents 174 classes

Goyal et al. The “something something” video database for learning and evaluating common sense. ICCV 2017.

Something-something

1. Category Selection

* Authors created 175 something-something templates

e.g.,

10 selected classes
Dropping [something]
Moving [something] from right to left
Moving [something] from left to right
Picking [something] up
Putting [something]
Poking [something]
Tearing [something]
Pouring [something]
Holding [something]
Showing [something] (almost no hand)

Goyal et al. The “something something” video database for learning and evaluating common sense. ICCV 2017.

Something-something

1. Category Selection

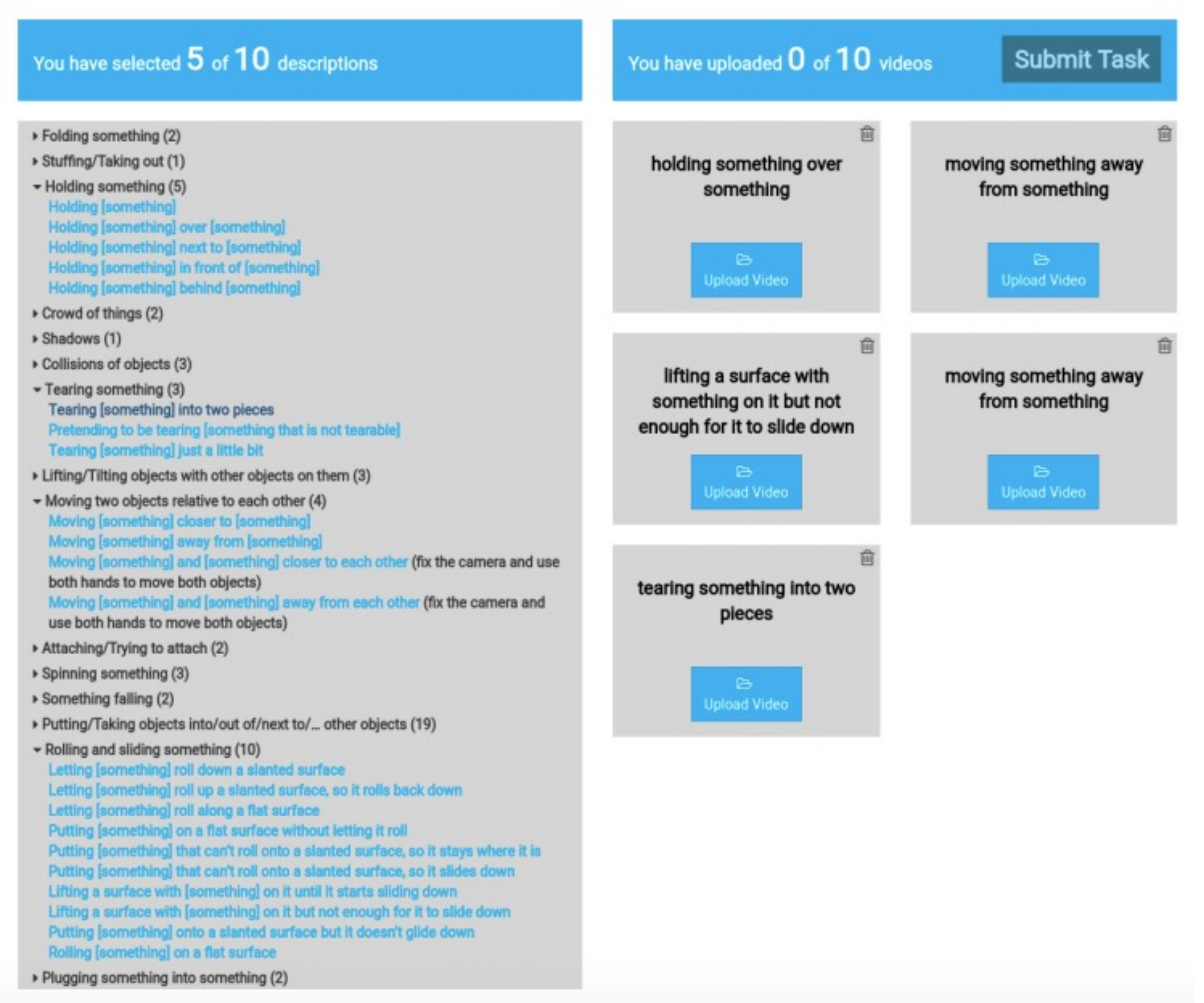
* Authors created 175 something-something templates

2. Video Collection

* Crowd workers submitted videos of them recording an implementation of the template

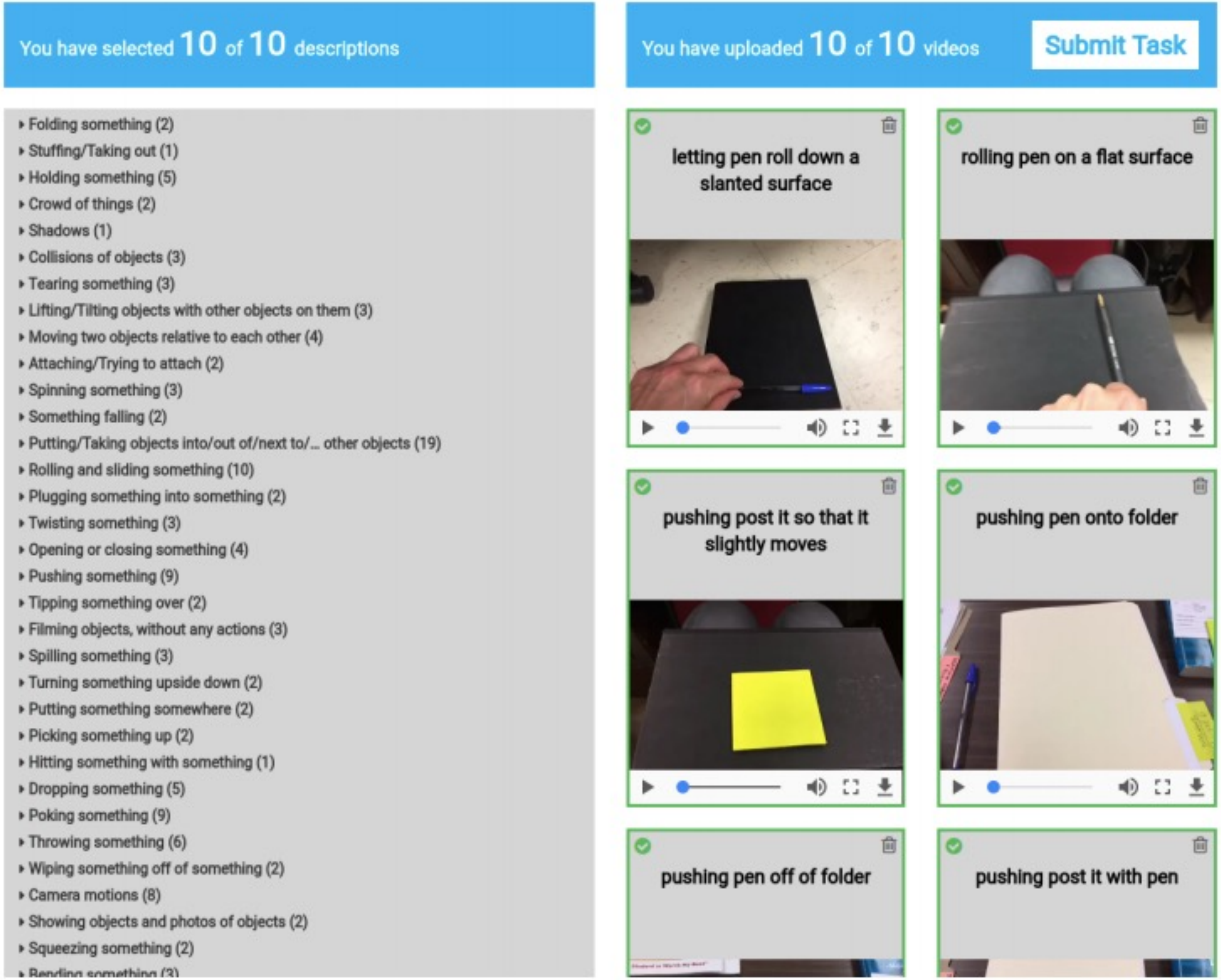
Goyal et al. The “something something” video database for learning and evaluating common sense. ICCV 2017.

UI



Goyal et al. The “something something” video database for learning and evaluating common sense. ICCV 2017.

UI



Goyal et al. The “something something” video database for learning and evaluating common sense. ICCV 2017.

Something-something Challenge

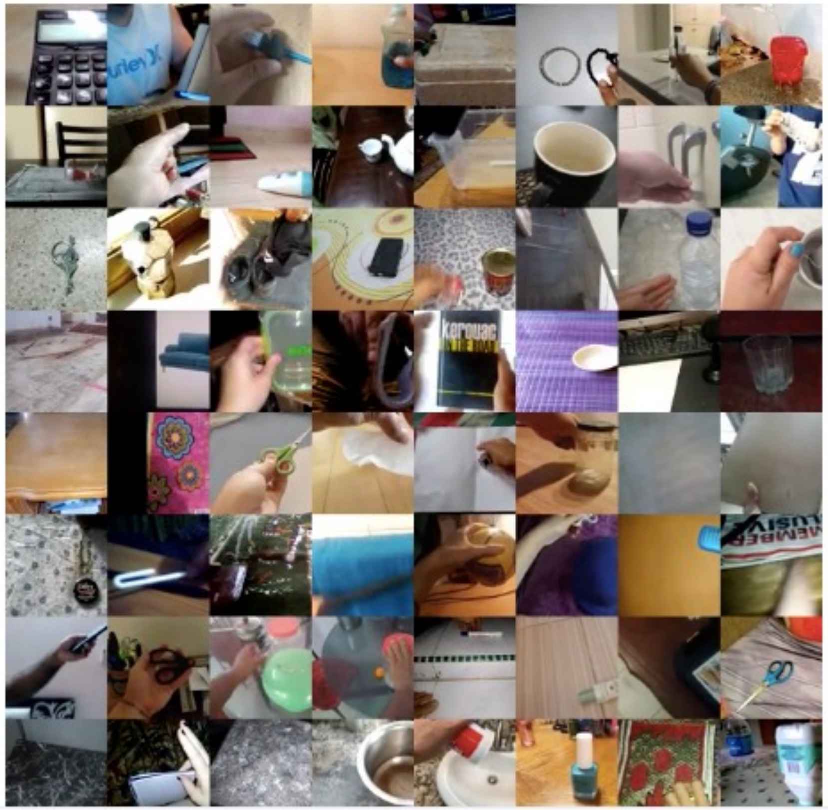
The 20BN-something-something Dataset V2

Introduction

The 20BN-SOMETHING-SOMETHING dataset is a large collection of densely-labeled video clips that show **humans performing pre-defined basic actions with everyday objects**. The dataset was created by a large number of crowd workers. It allows machine learning models to develop fine-grained understanding of basic actions that occur in the physical world. It is **available free of charge for academic research**. Commercial licenses are available upon request.

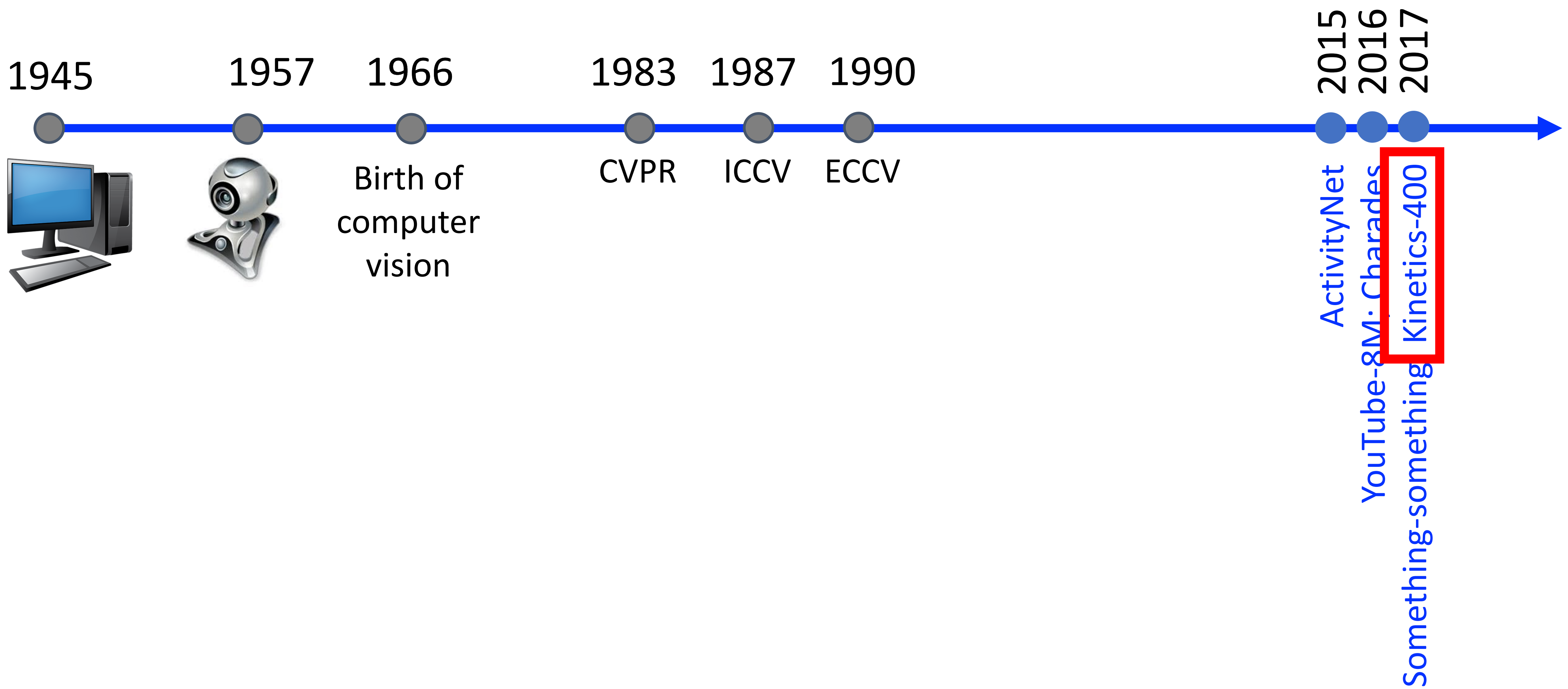
This is the second release of the dataset. The first release is also still available here. The new release features the following updates:

- **Greatly increased number of videos:** With 220,847 videos (vs. 108.499 in V1) we release more than twice as many videos.
- **Object annotations and captioning:** For each video in the training and validation sets we now also provide object annotations in addition to the video label if applicable. For example, for a label like "Putting



https://20bn.com/datasets/something-something

Datasets


1945
1957
1966
1983
1987
1990
2015
2016
2017
Birth of computer vision
CVPR
ICCV
ECCV
ActivityNet
YouTube-8M; Charades
Kinetics-400
Something-something

Kinetics-400

Multi-class classification dataset of human actions videos that is two orders of magnitude larger than prior work

- 306,245 videos that each roughly 10 seconds long
- Represents 400 classes covering with 400-1150 clips per class:
  - Person Actions; e.g., drawing, drinking, laughing, punching
  - Person-Person Actions; e.g., hugging, kissing, shaking hands;
  - Person-Object Actions; e.g., opening presents, mowing lawn, washing dishes

Will Kay al. The Kinetics Human Action Video Dataset. arXiv 2017.

Kinetics-400

1. Category Selection

* Categories taken from existing datasets and AMT workers feedback about more suitable categories for clips with existing category labels assigned

→

2. Video Collection

* Image action recognition classifiers applied to identify relevant clips (5 seconds before plus 5 seconds after image where activity is recognized)

→

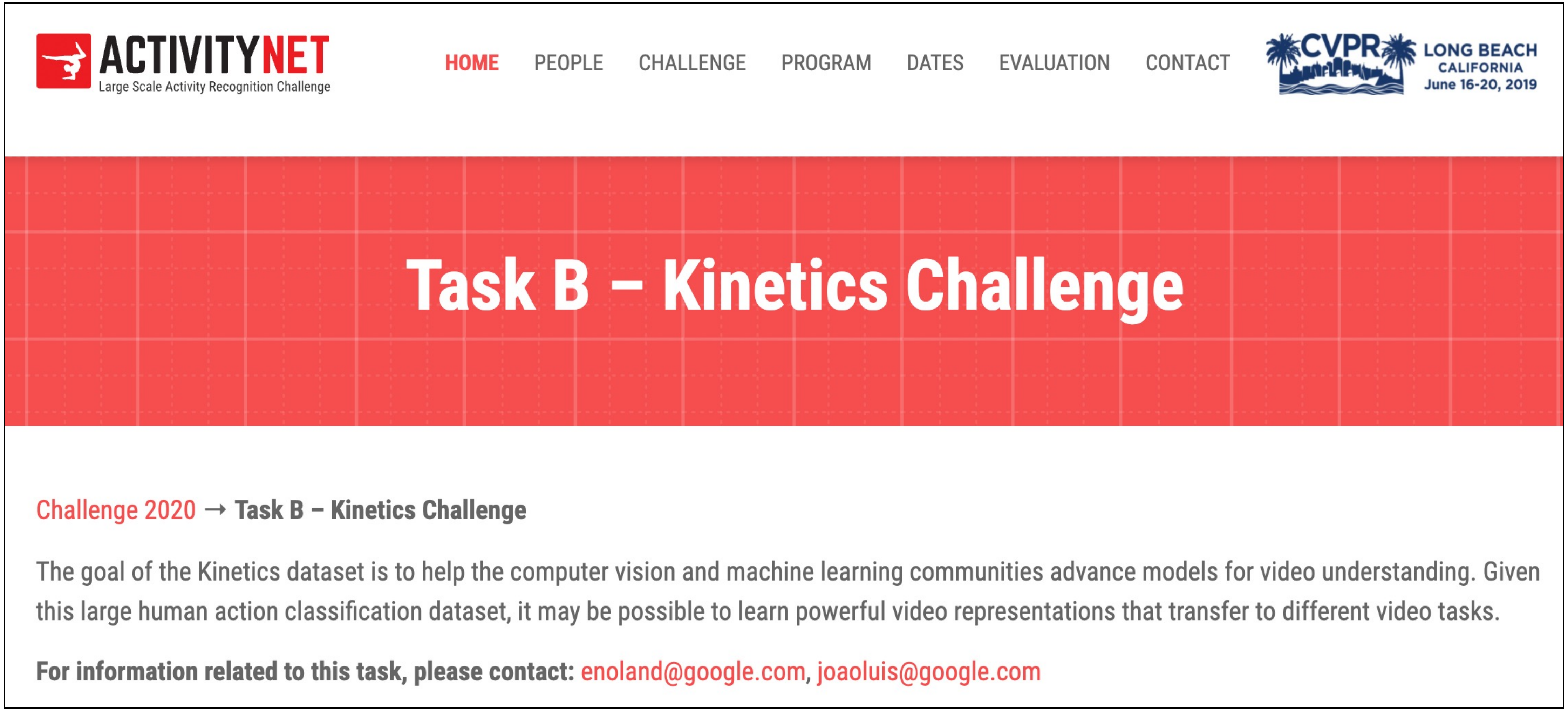
3. Video Verification

* AMT worker verifies action is present in each

* Label determined by majority of 5 workers

Will Kay al. The Kinetics Human Action Video Dataset. arXiv 2017.

Kinetics Challenge & Annual Workshop


ACTIVITYNET
Large Scale Activity Recognition Challenge

HOME PEOPLE CHALLENGE PROGRAM DATES EVALUATION CONTACT

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Task B – Kinetics Challenge

Challenge 2020 → **Task B – Kinetics Challenge**

The goal of the Kinetics dataset is to help the computer vision and machine learning communities advance models for video understanding. Given this large human action classification dataset, it may be possible to learn powerful video representations that transfer to different video tasks.

For information related to this task, please contact: enoland@google.com, joaoluis@google.com

http://activity-net.org/challenges/2020/tasks/guest_kinetics.html

Class Task: Video Classification Costs

Assume the task is to classify the presence of 10 activities in 1,000,000 30-second videos. **How much do you believe it will cost in US dollars to collect all the crowdsourced annotations for the datasets?**

Video Classification: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models

Accuracy Metric

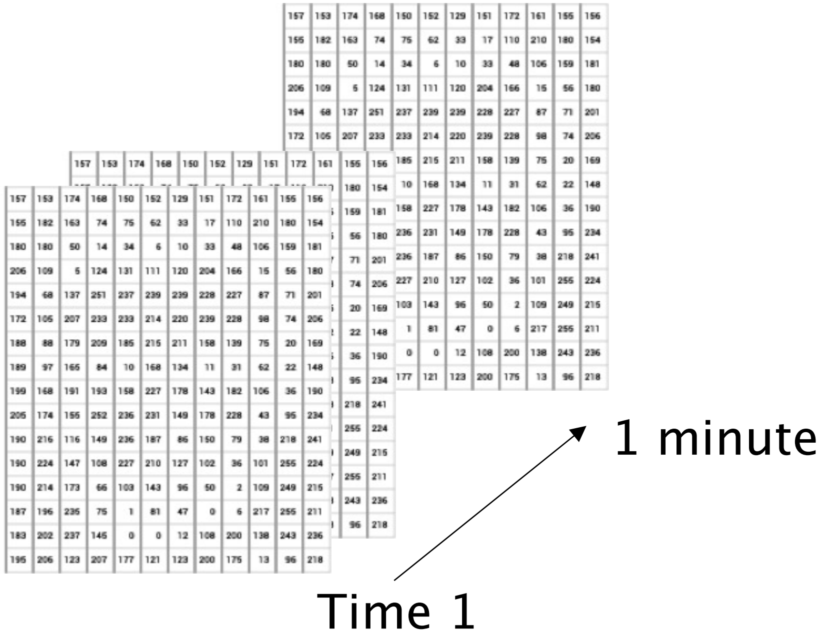
- Percentage of correct predictions

Video Classification: Today's Topics

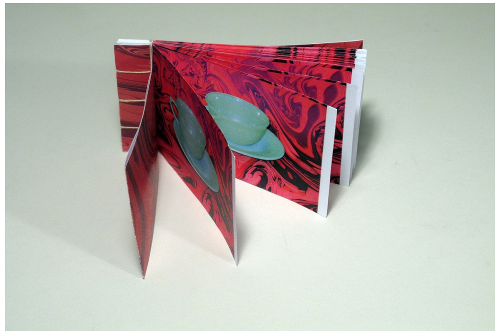
- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models

Key Idea

- Recall: a video is a series of images



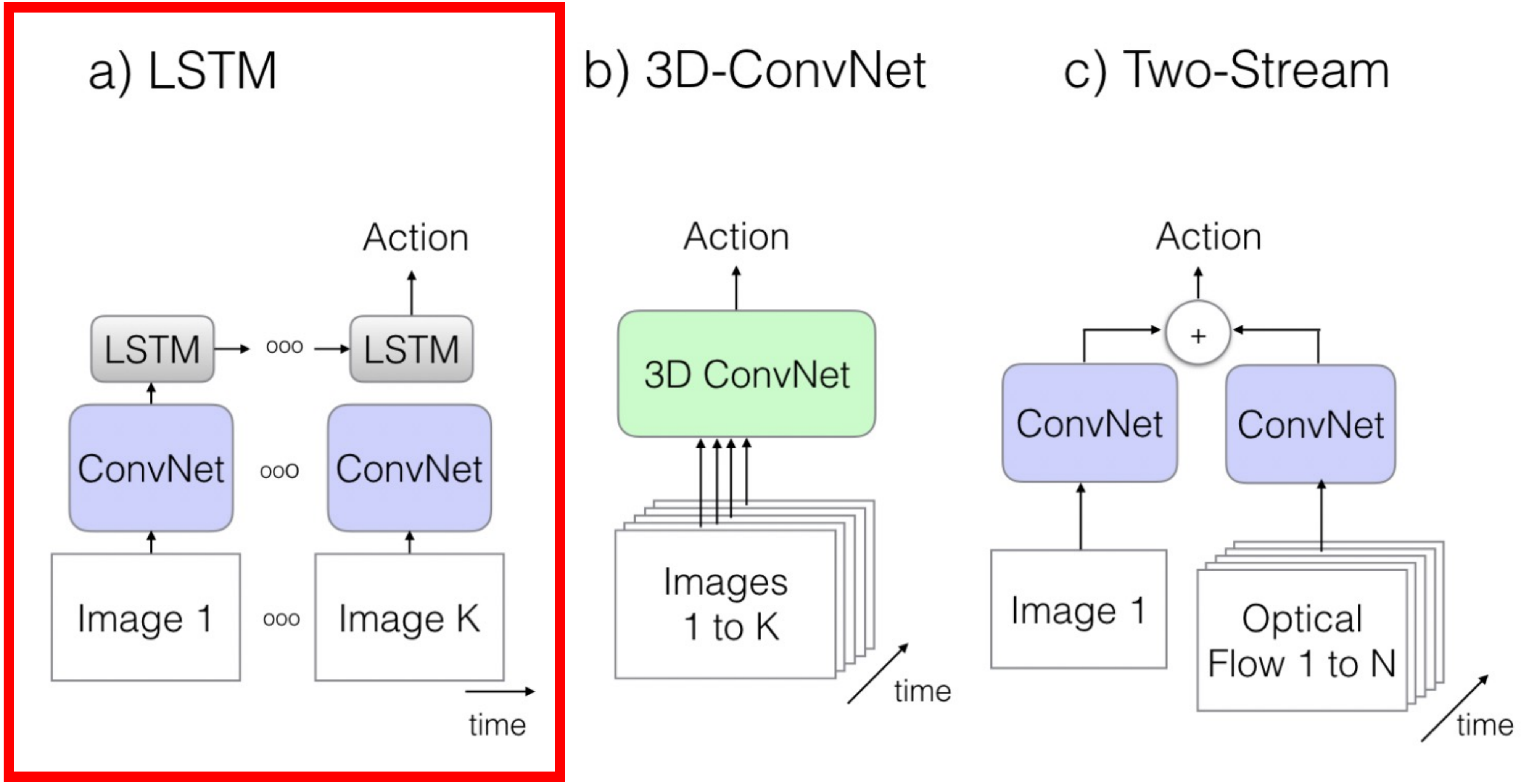
Analogous to:



- How to go beyond image-based techniques and improve techniques by considering the temporal relationship between frames in a video? (e.g., recognize difference between a door that is opening vs closing)

Approaches to Capture Temporal Information

LSTM is a type of recurrent neural network (rnn); more on this Wednesday!



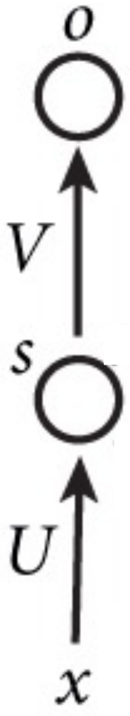
Source: Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017.

Recurrent Neural Networks (RNNs)

- Main idea: use hidden state to capture information about the past

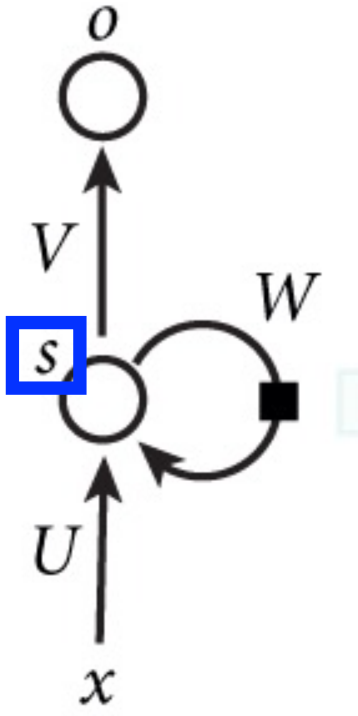
Feedforward Network

Each layer receives input from the previous layer with no loops



Recurrent Network

Each layer receives input from the previous layer and the output from the previous time step



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

Recurrent Neural Networks (RNNs)

- Main idea: use hidden state to capture information about the past

Recurrence formula applied at every time step:

Model parameters

$$\boxed{s_t} = \boxed{f_m}\boxed{s_{t-1}}\boxed{x_t}$$

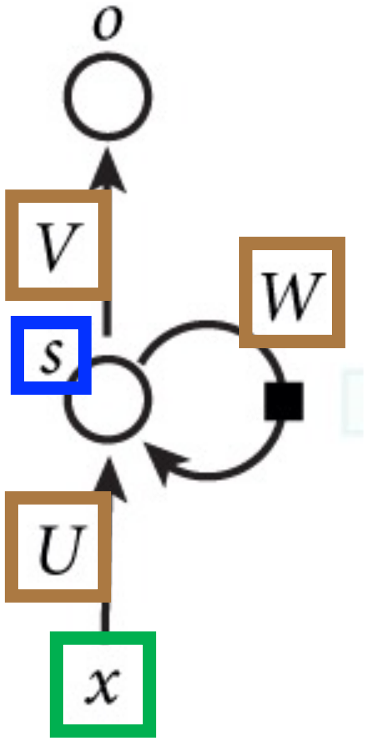
New state

Old state

Input at time step

Recurrent Network

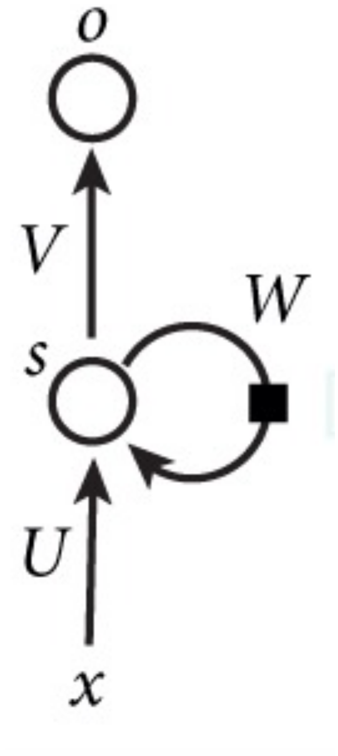
Each layer receives input from the previous layer and the output from the previous time step



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Time Step 1

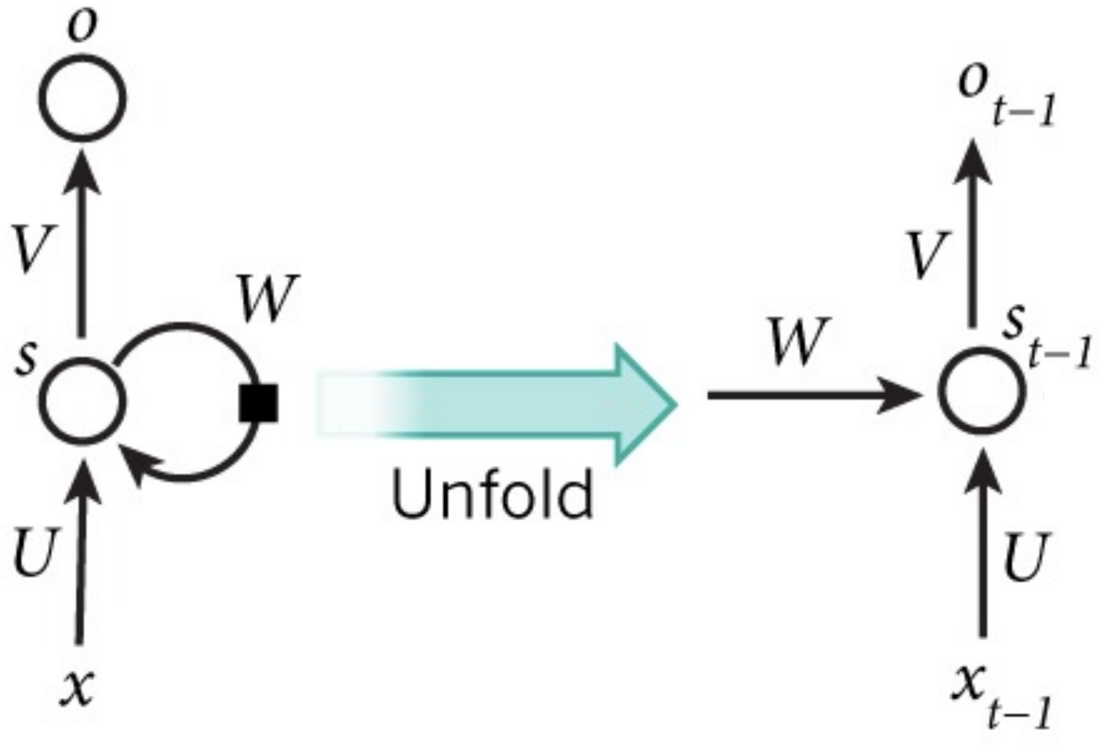
- Main idea: use hidden state to capture information about the past



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Time Step 1

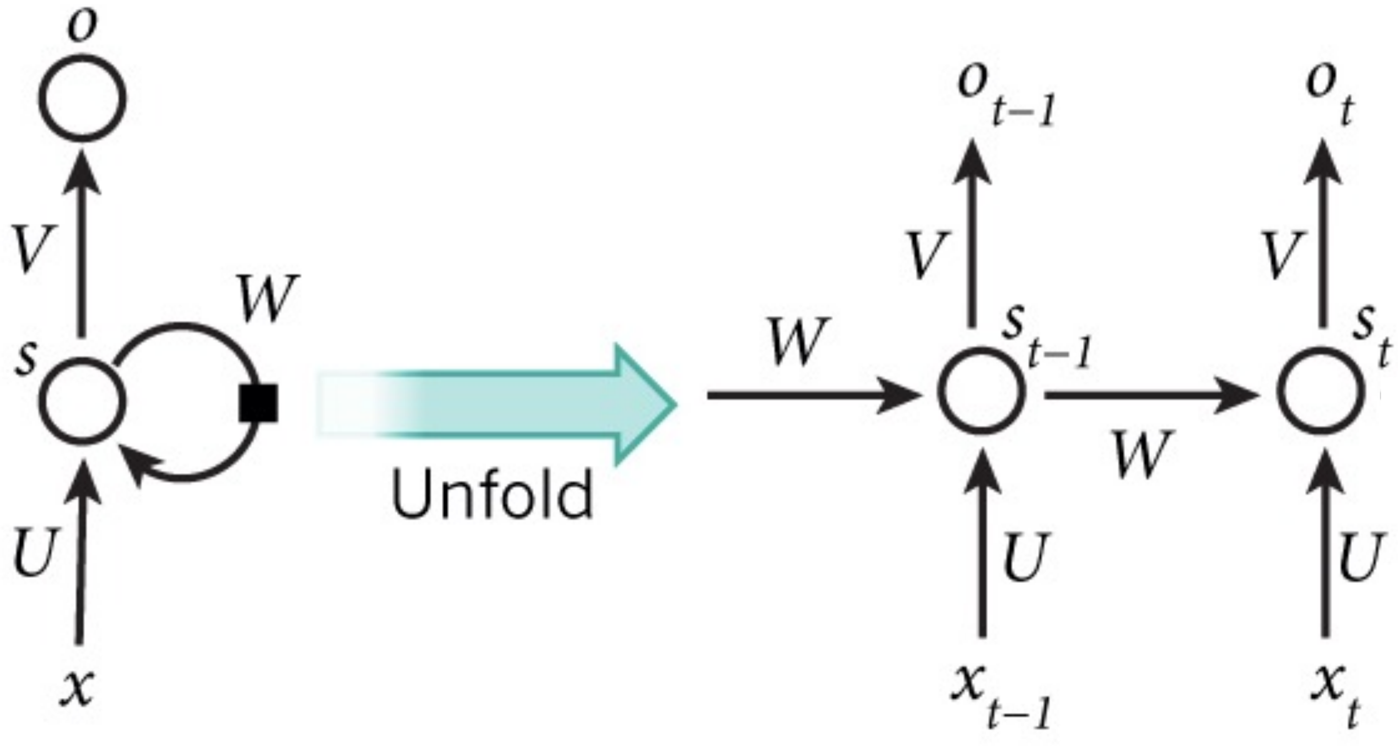
- Main idea: use hidden state to capture information about the past



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Time Step 2

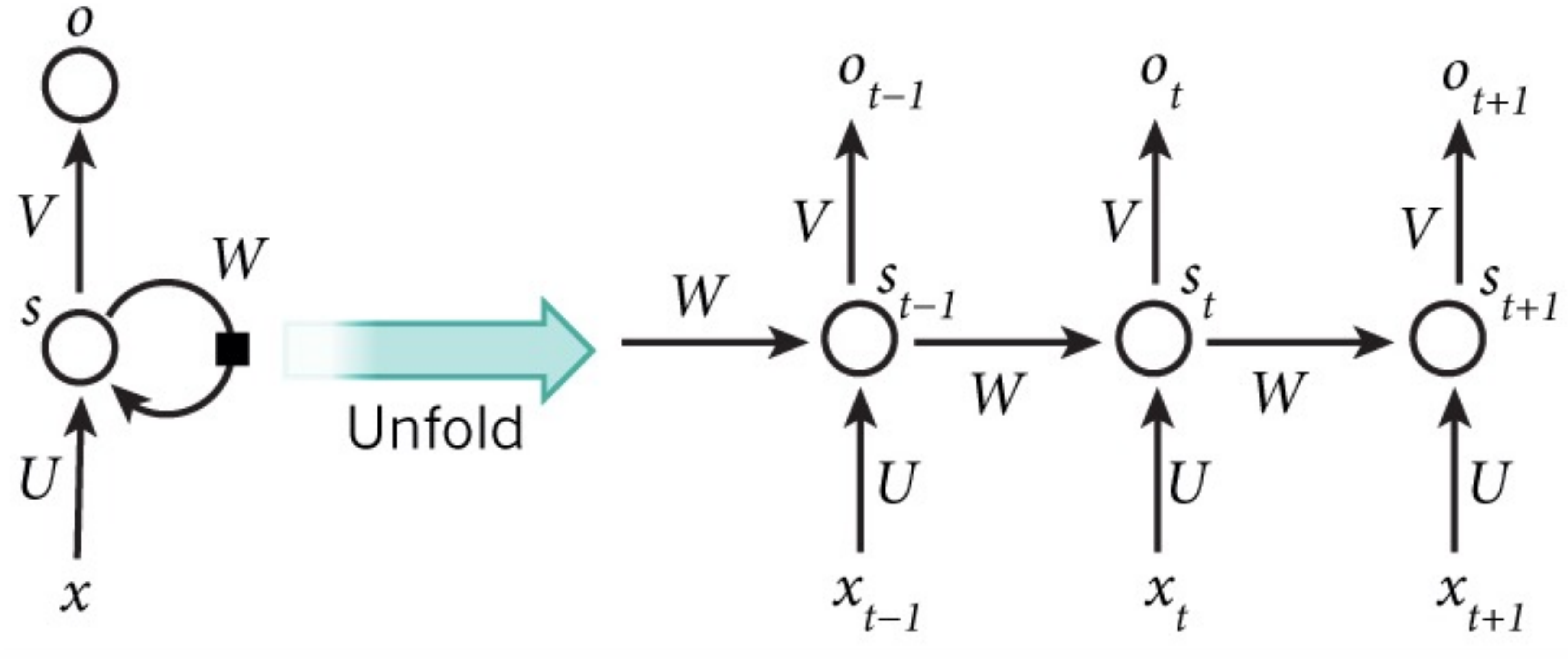
- Main idea: use hidden state to capture information about the past



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Time Step 3

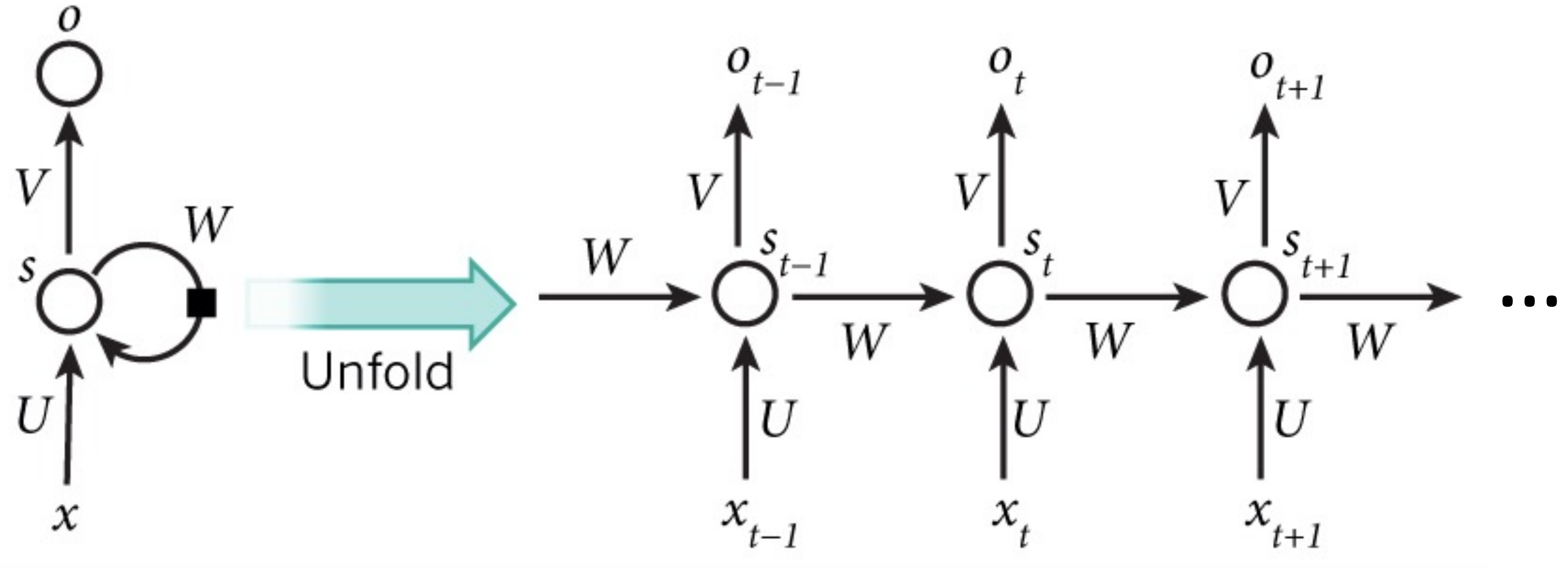
- Main idea: use hidden state to capture information about the past



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: And So On…

- Main idea: use hidden state to capture information about the past



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Model Parameters and Inputs

- Main idea: use hidden state to capture information about the past

Recurrence formula applied at every time step:

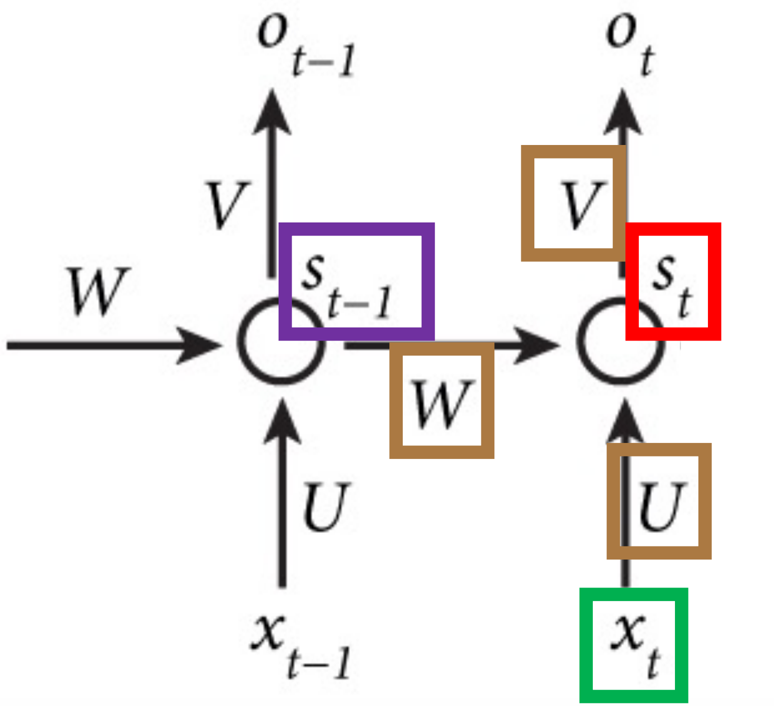
Model parameters

$$\boxed{s_t} = \boxed{f_m}(\boxed{s_{t-1}}, \boxed{x_t})$$

New state

Old state

Input at time step



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Model Parameters and Inputs

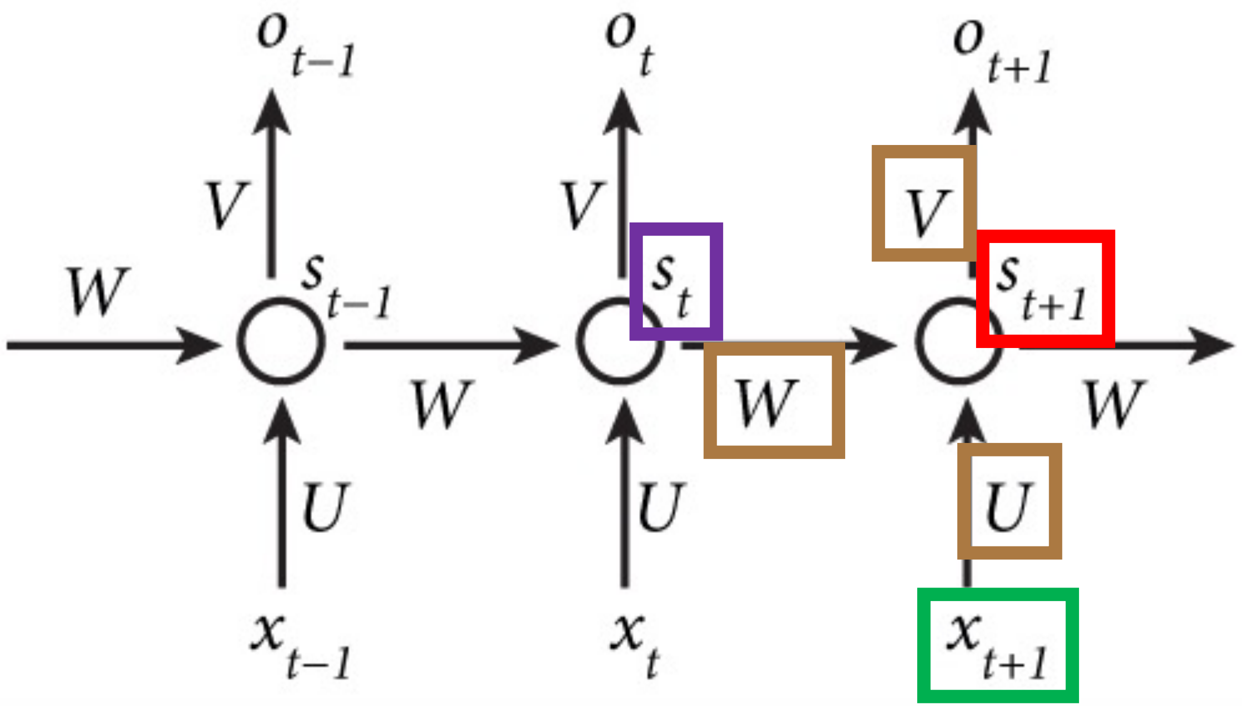
- Main idea: use hidden state to capture information about the past

Recurrence formula applied at every time step:

Model parameters

$$\boxed{s_t} = \boxed{f_m}(\boxed{s_{t-1}}, \boxed{x_t})$$

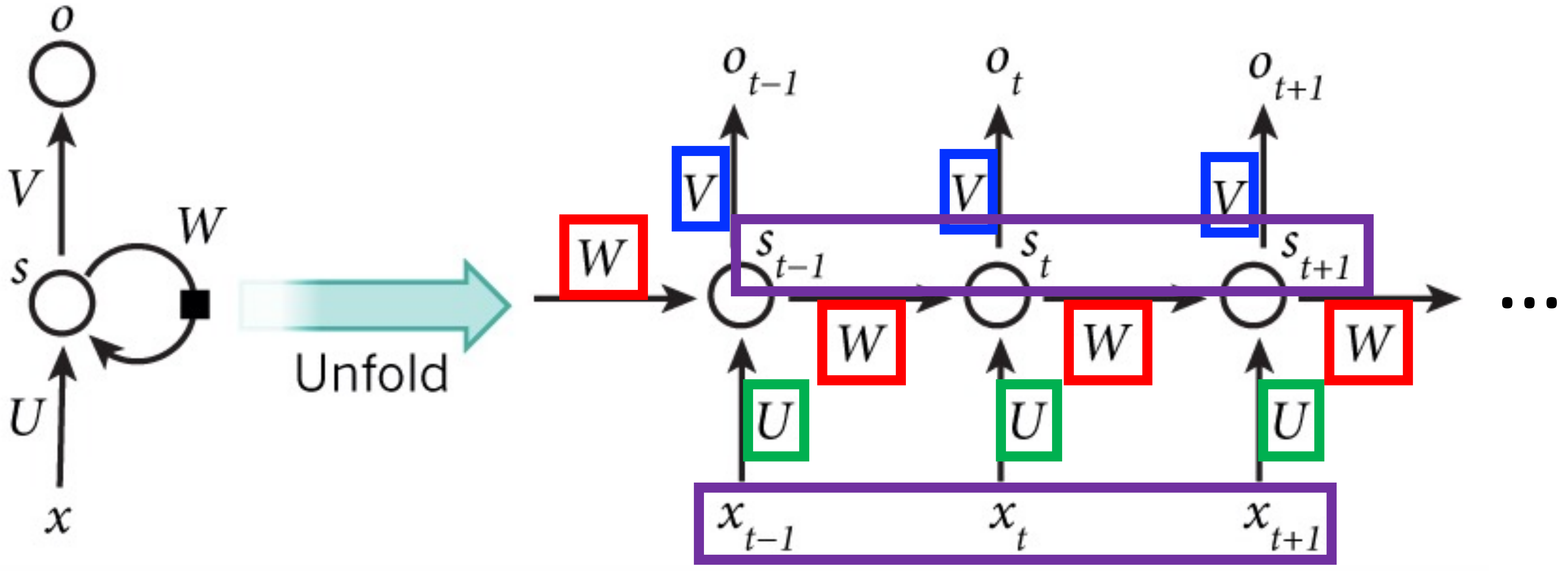
New state | Old state | Input at time step



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Model Parameters and Inputs

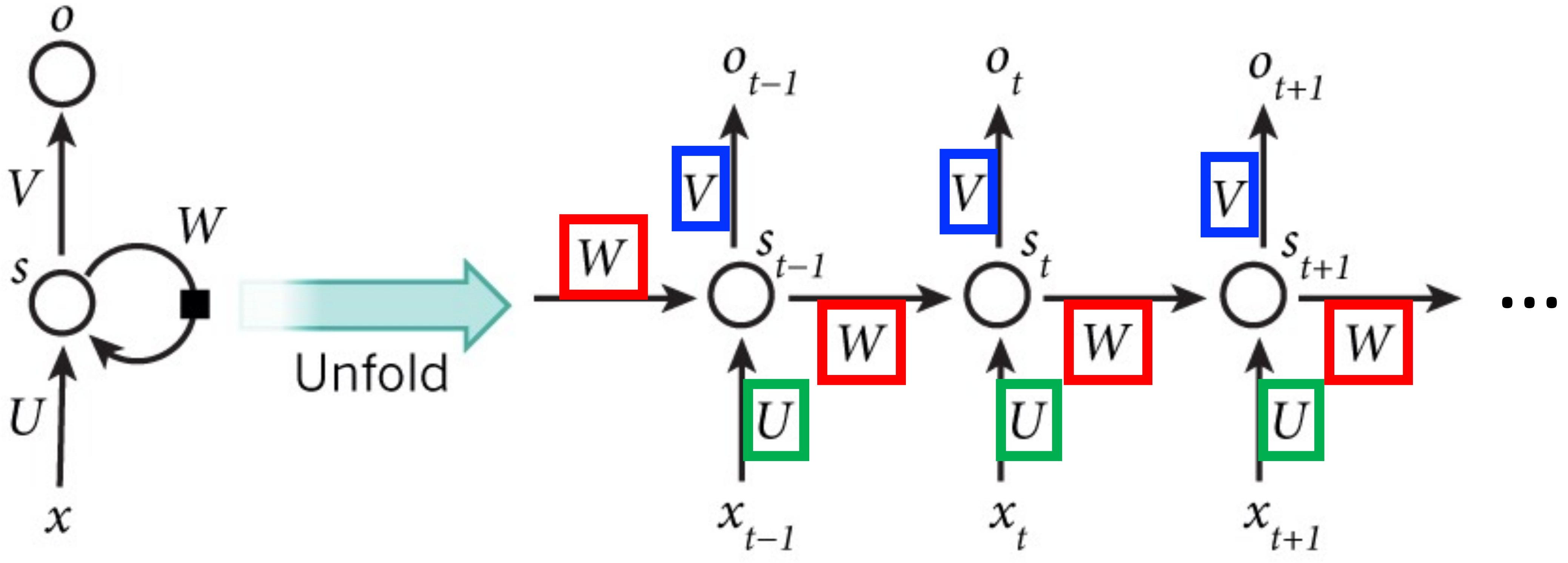
- All layers share the same model parameters (U, V, W)
  - What is different between the layers?



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

RNN: Model Parameters and Inputs

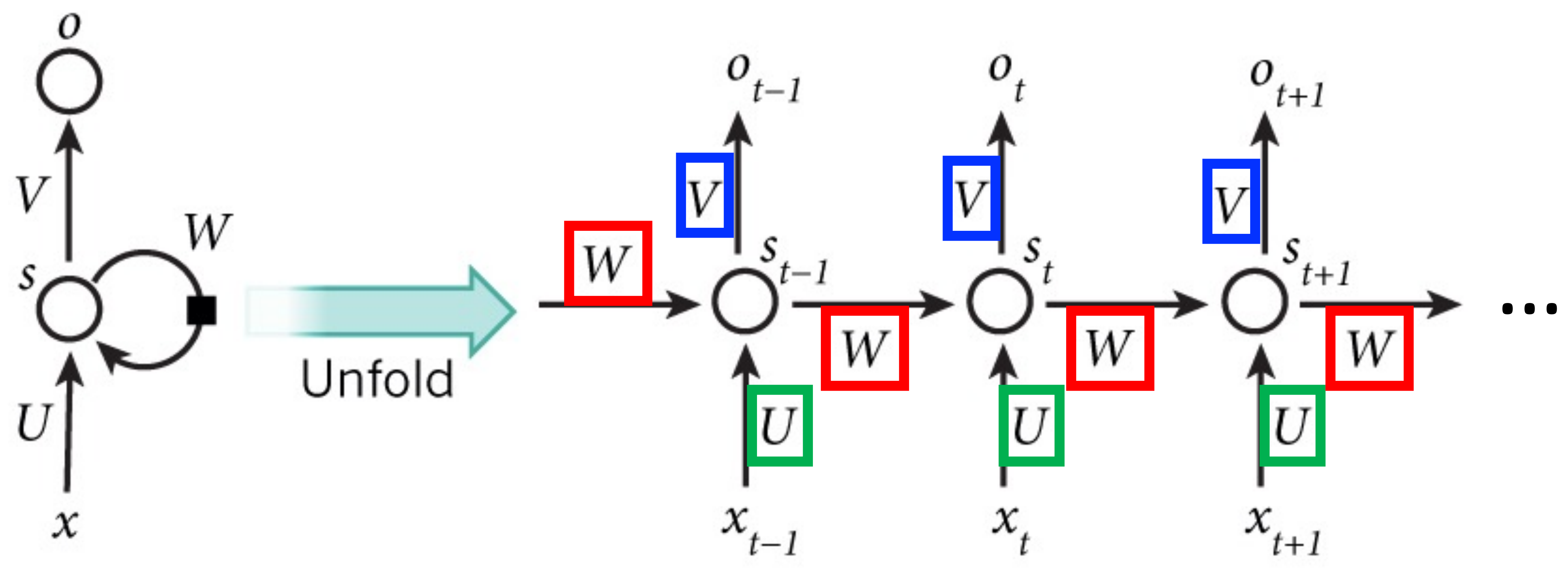
- When unfolded, a RNN is a deep feedforward network with shared weights!



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

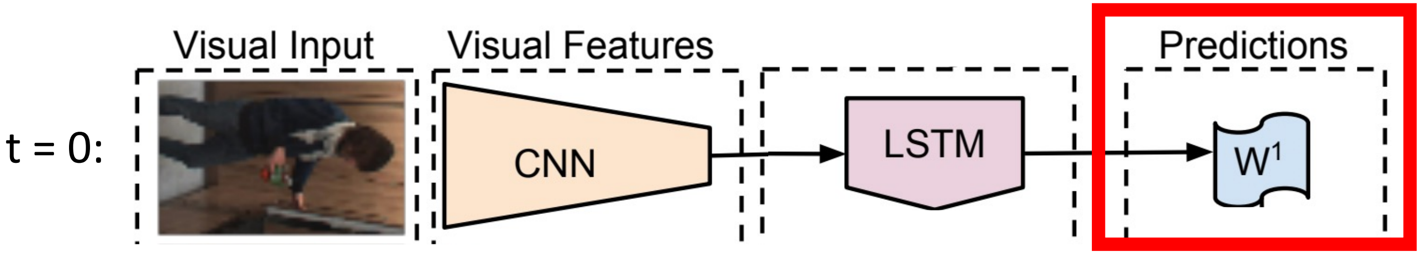
RNN: Advantage

- Retains information about past inputs for an amount of time that depends on the model's weights and input data rather than a fixed duration selected a priori



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

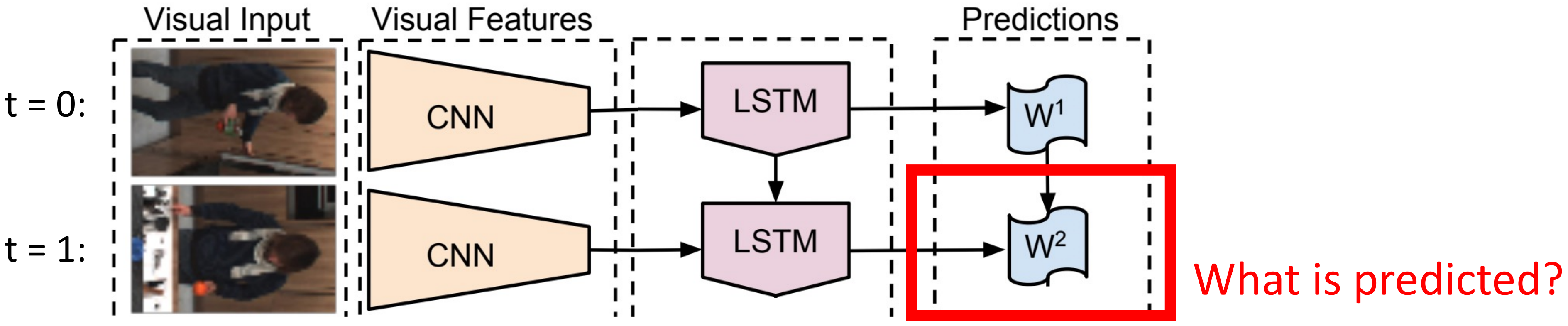
RNN for Video Classification



What is predicted?

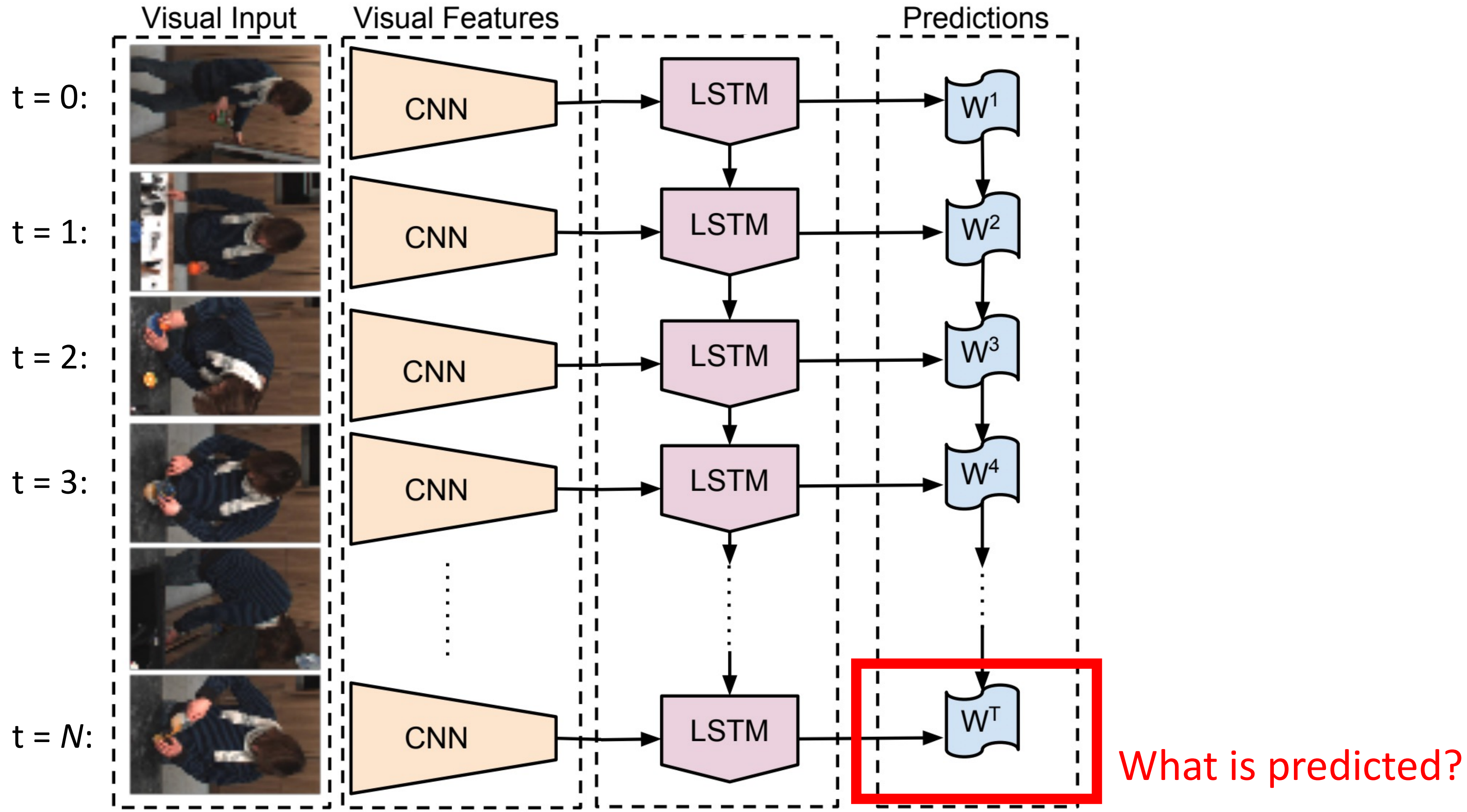
Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

RNN for Video Classification



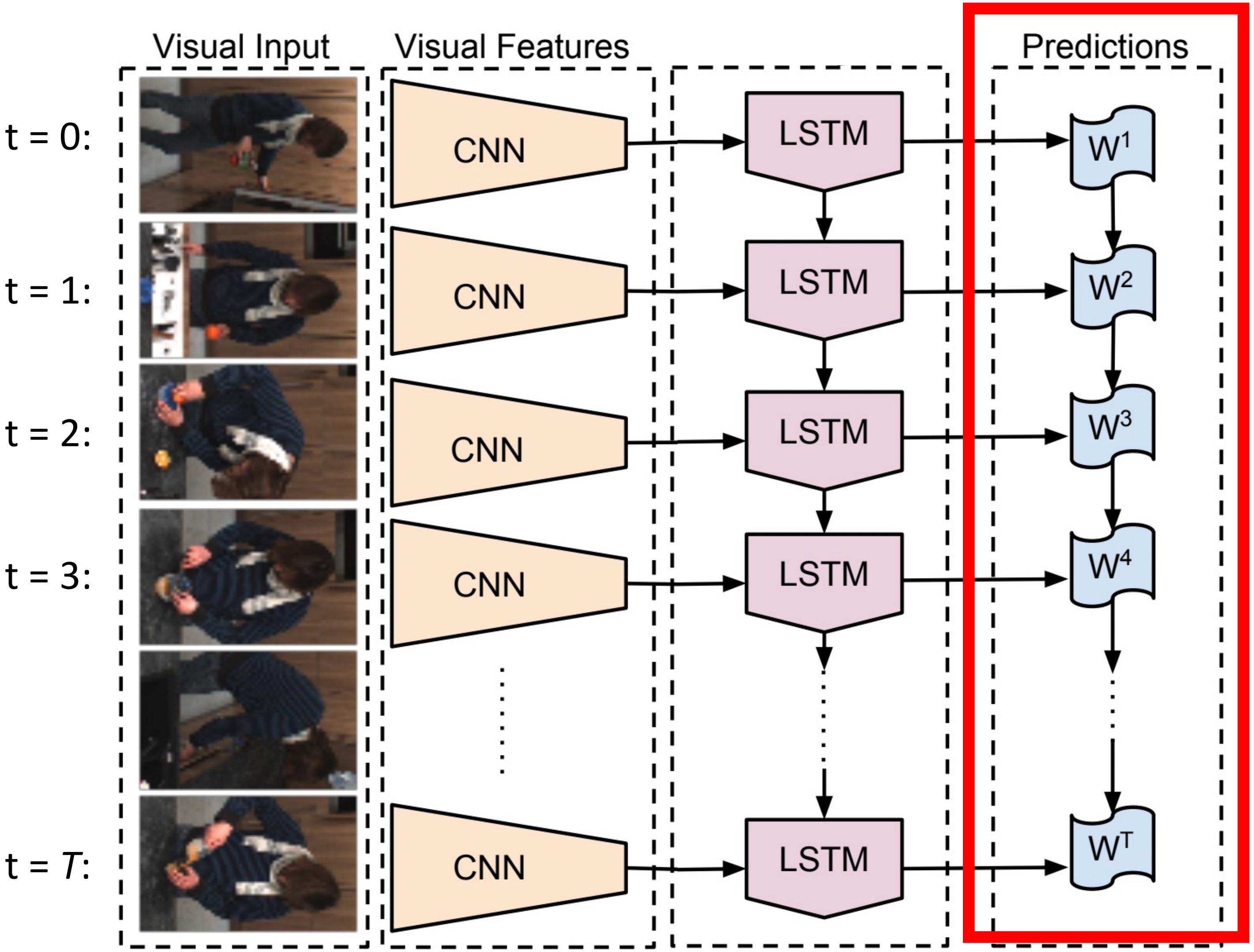
Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

RNN for Video Classification



Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

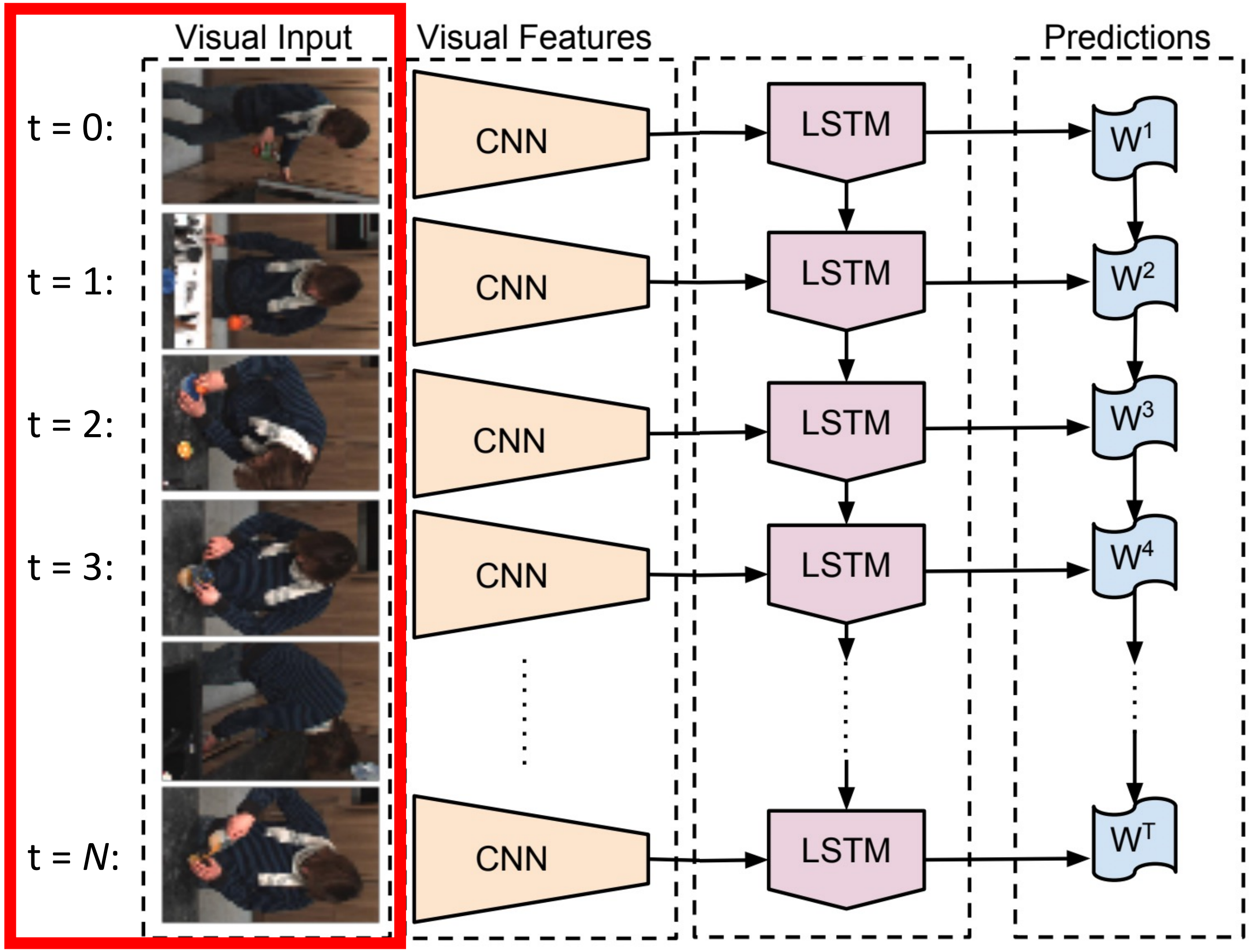
RNN for Video Classification



How should the final classification prediction be made (recall predictions are made from time step 1 to time step T)?

Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

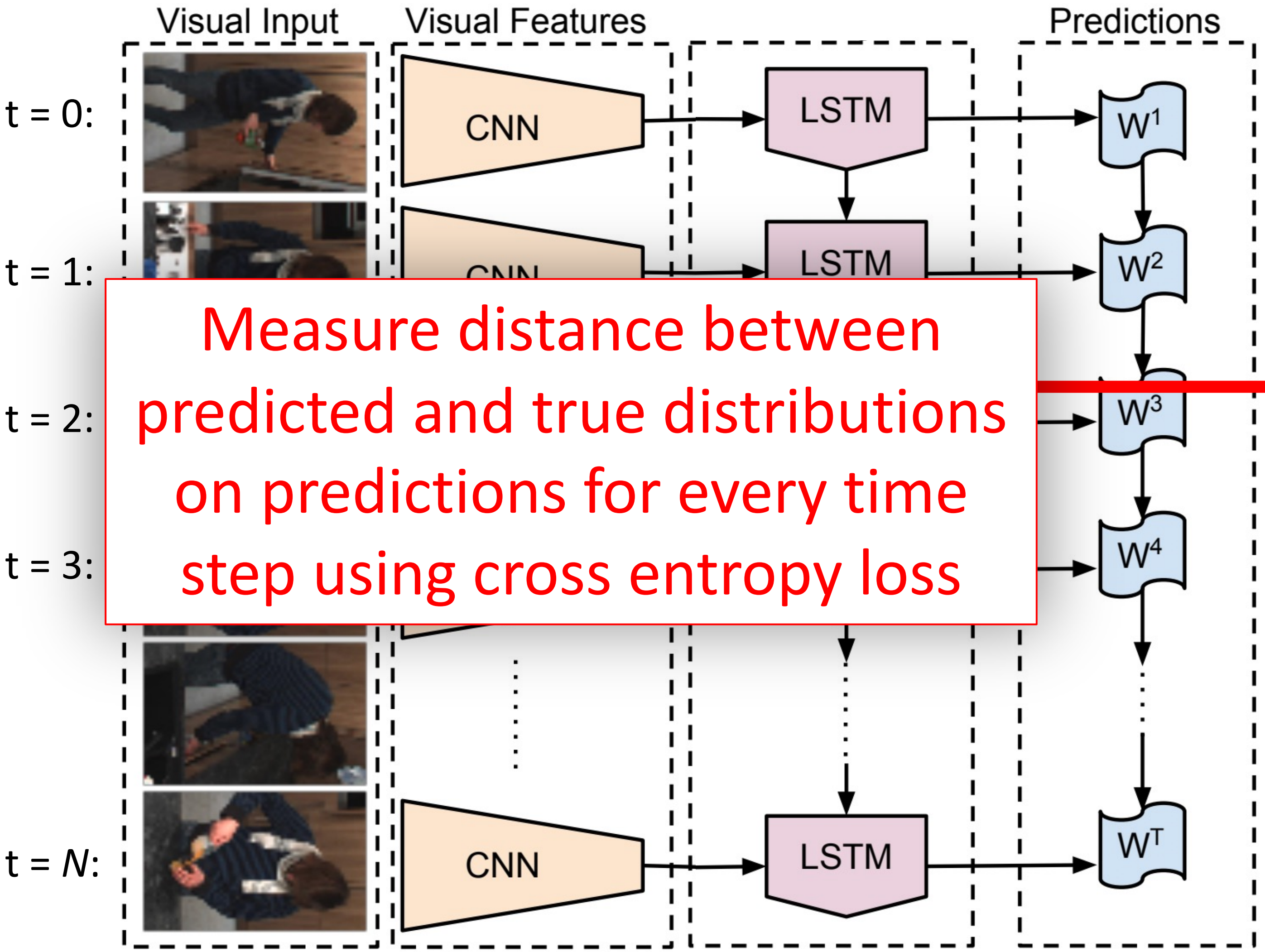
RNN for Video Classification



What input duration is supported?

Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

RNN for Video Classification: Training Algorithm

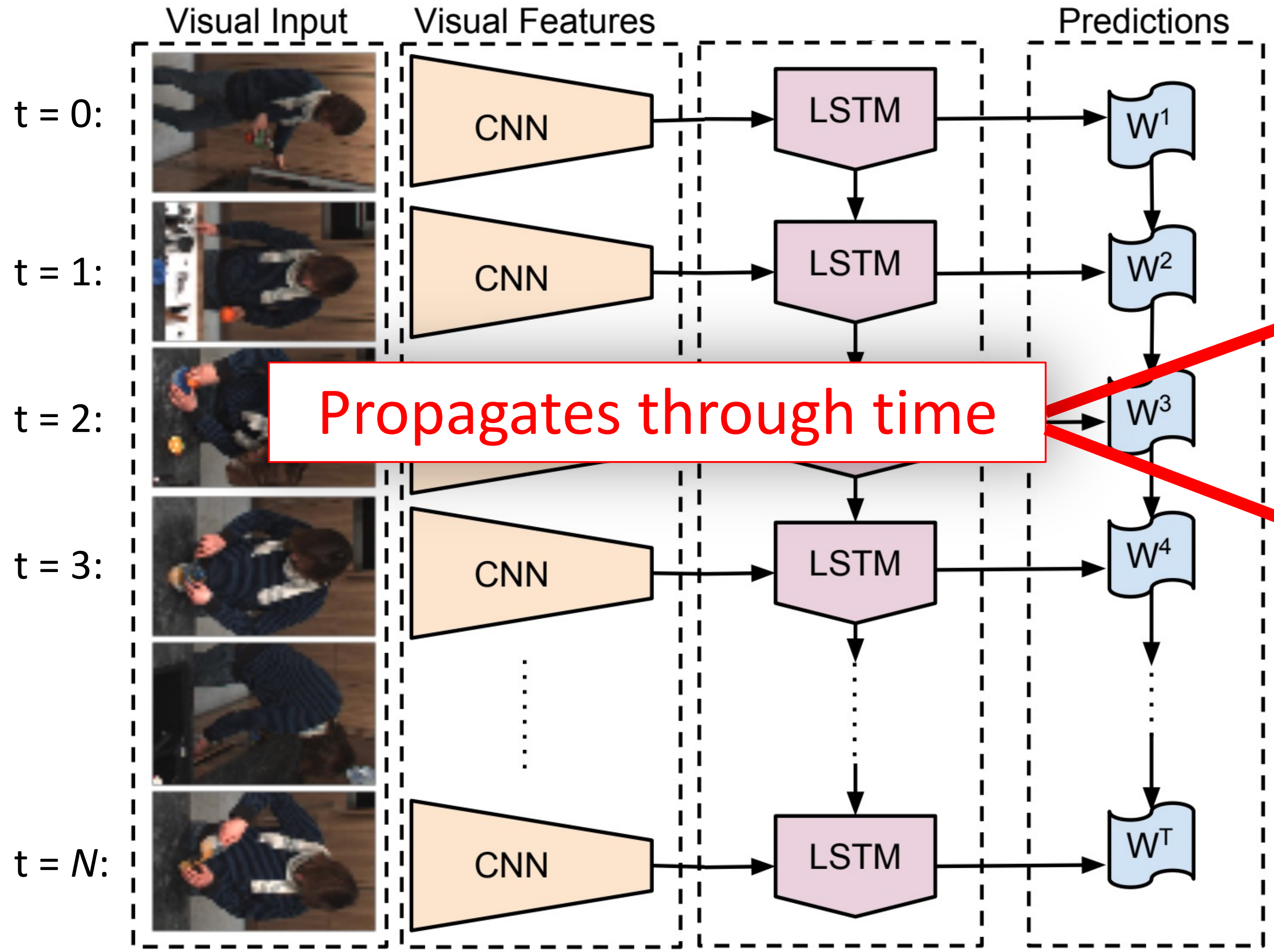


Measure distance between predicted and true distributions on predictions for every time step using cross entropy loss

- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

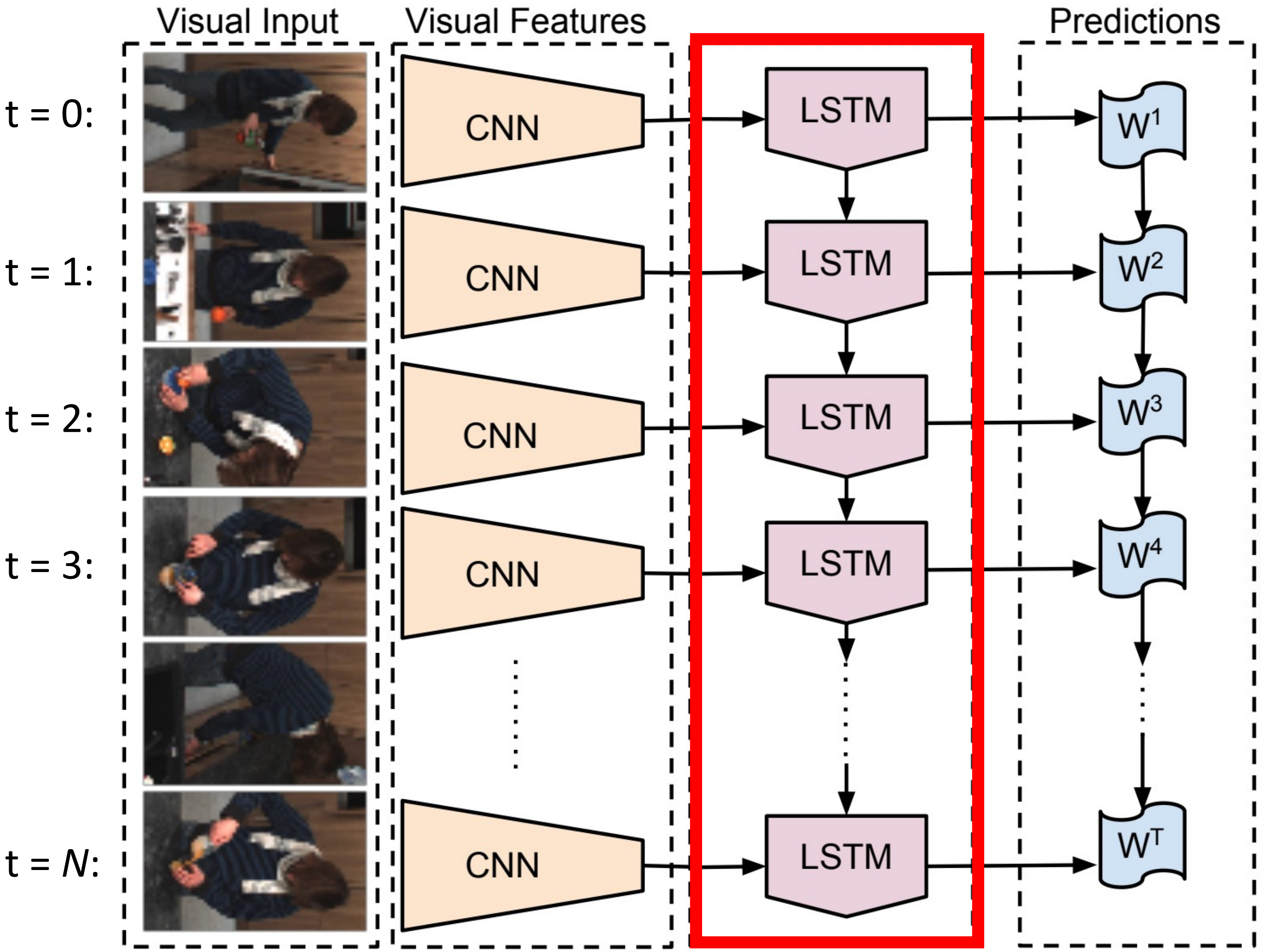
RNN for Video Classification: Training Algorithm



- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

RNN for Video Classification



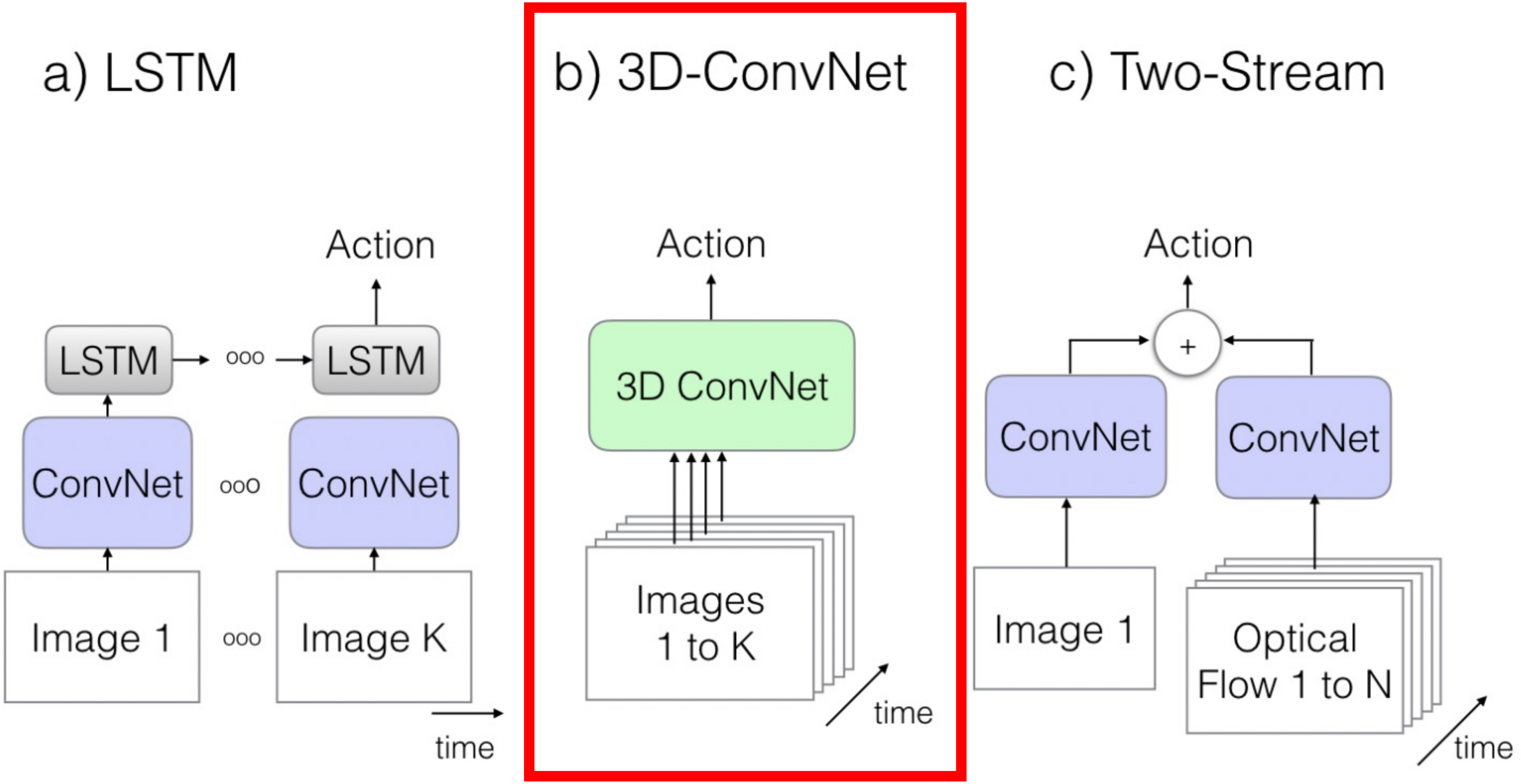
Recall: The LSTM layer's weights and input data determine what information about the past gets propagated to later time steps

Source: Jeff Donahue et al. Long-term recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015.

RNN for Video Classification: Limitations

- Successful training requires many videos which makes RNNs resource-hungry and time-consuming

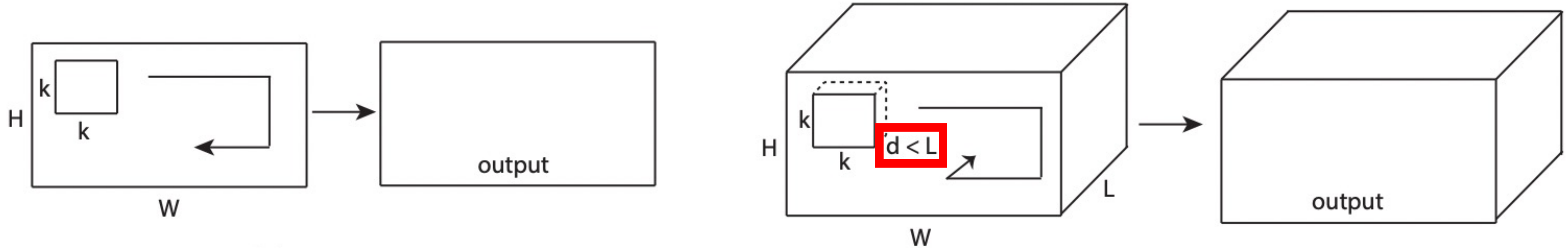
Approaches to Capture Temporal Information



Source: Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017.

ConvNet Architecture

- CNNs with 3D kernels instead of 2D kernels to preserve temporal information in addition to spatial information

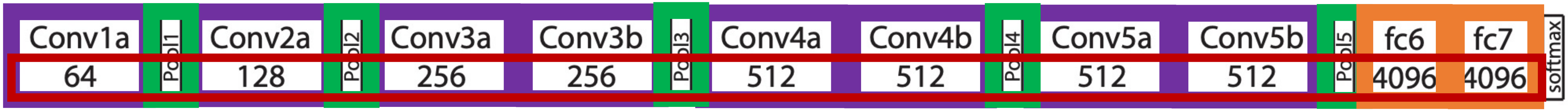


(a) 2D convolution (c) 3D convolution

Source: Du Tran et al. Learning Saptiotemporal Features with 3D Convolutional Networks. ICCV 2015.

ConvNet Architecture

- How many convolutional layers are there?
- How many pooling layers are there?
- How many fully-connected layers are there?

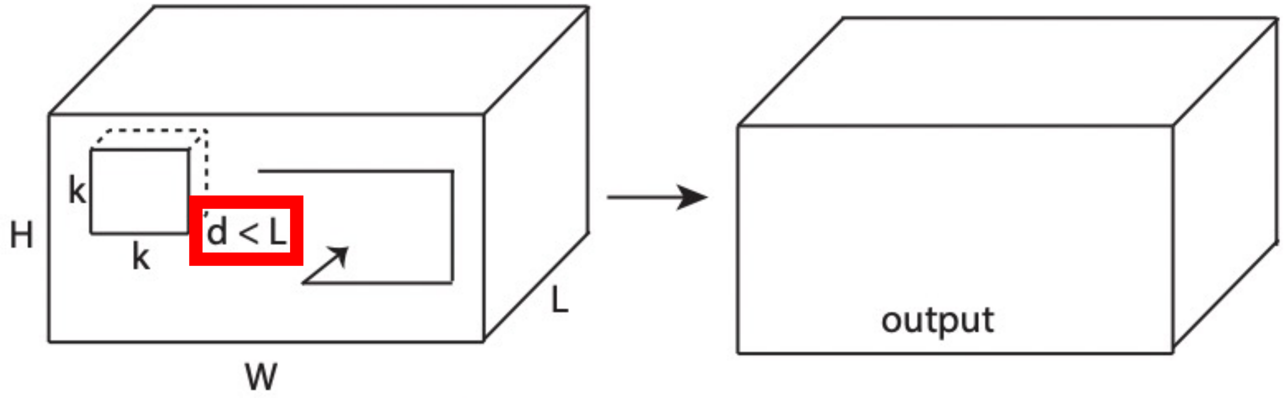


Numbers indicate number of kernels/nodes per layer

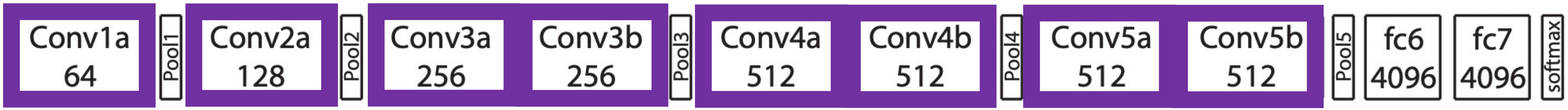
Source: Du Tran et al. Learning Saptiotemporal Features with 3D Convolutional Networks. ICCV 2015.

ConvNet Architecture

- Key question: what kernel depth should be used?
  - From experimentation: 3 at every layer (i.e., 3x3x3 kernels)



(c) 3D convolution



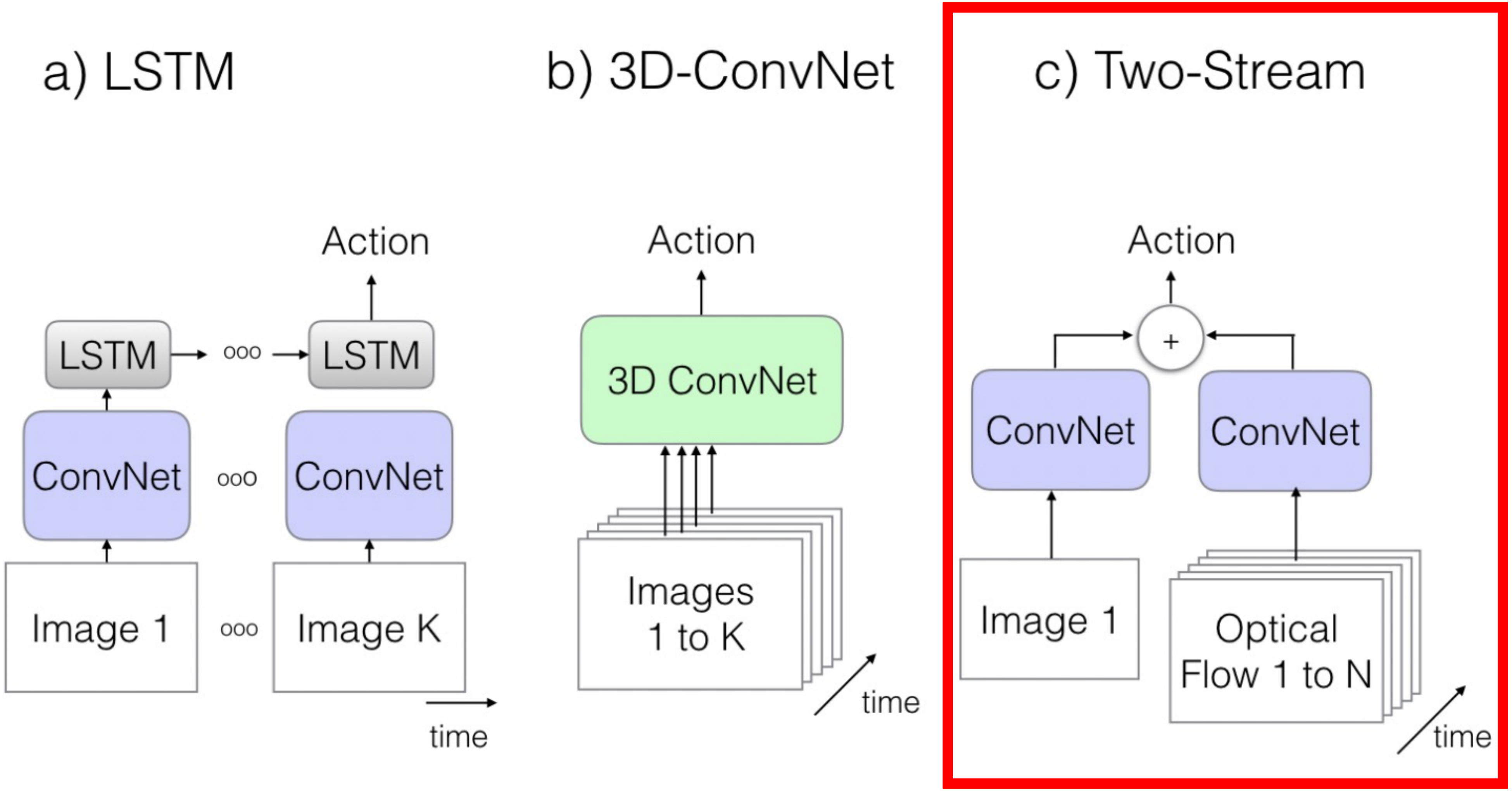
Source: Du Tran et al. Learning Saptiotemporal Features with 3D Convolutional Networks. ICCV 2015.

ConvNet: Limitations

- 3D kernels introduce many model parameters and so successful training requires many videos (i.e., resource-hungry and time-consuming)

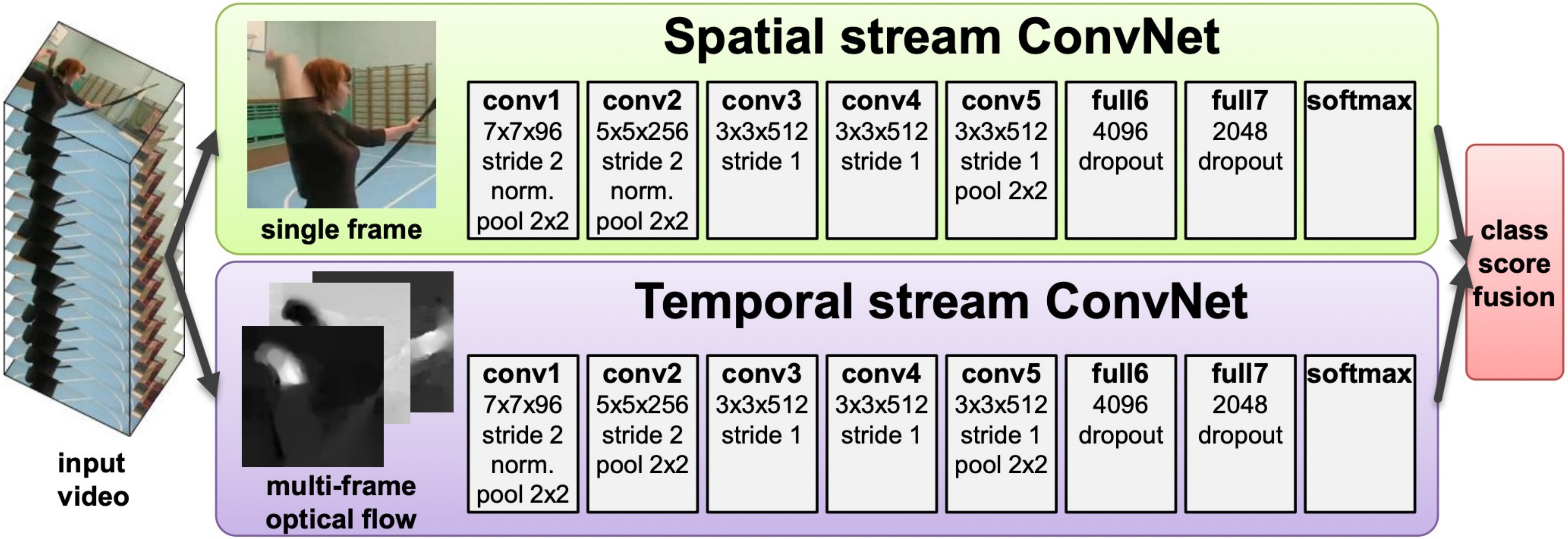
- Does not capture long-term temporal information (duration determined by depth of kernel)

Approaches to Capture Temporal Information



Source: Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017.

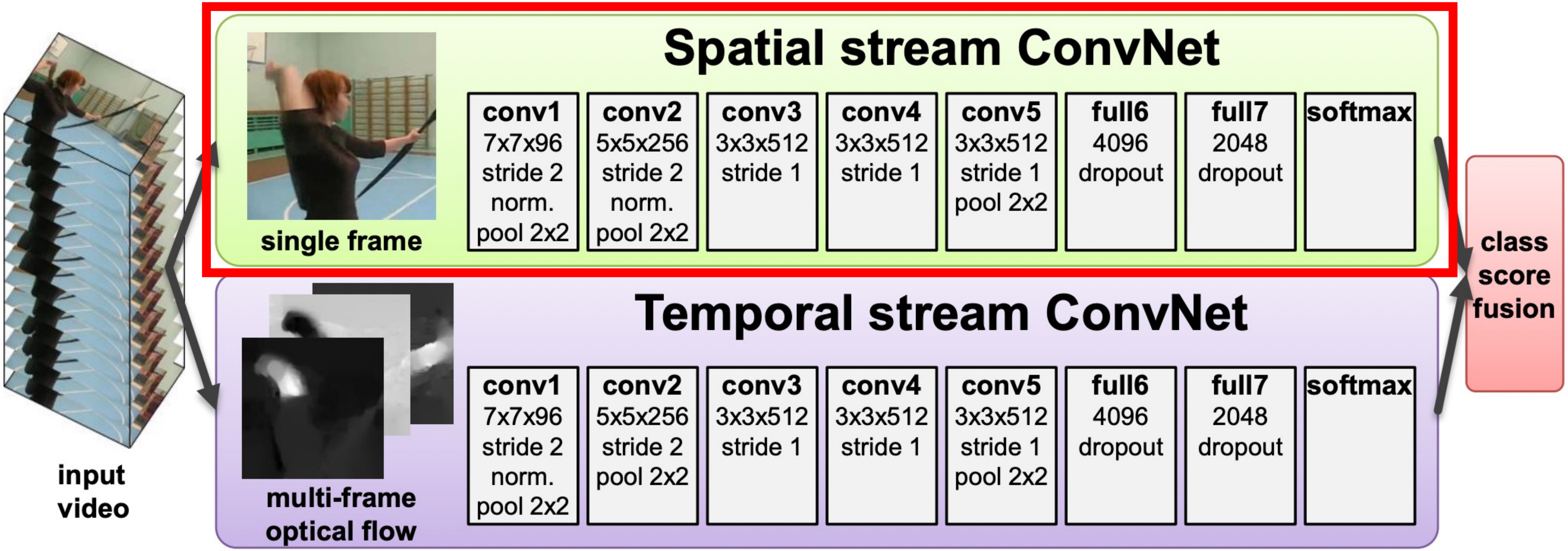
Two-Stream Architecture



Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture

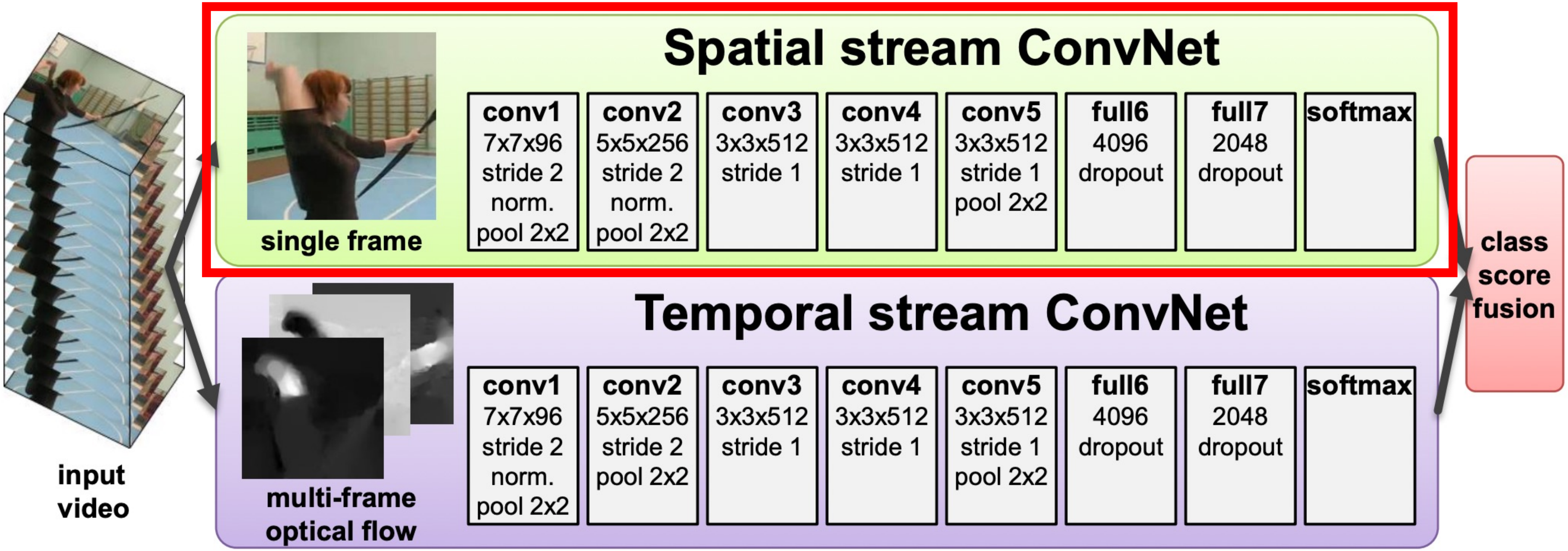
Learns to predict from still images the actions



Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

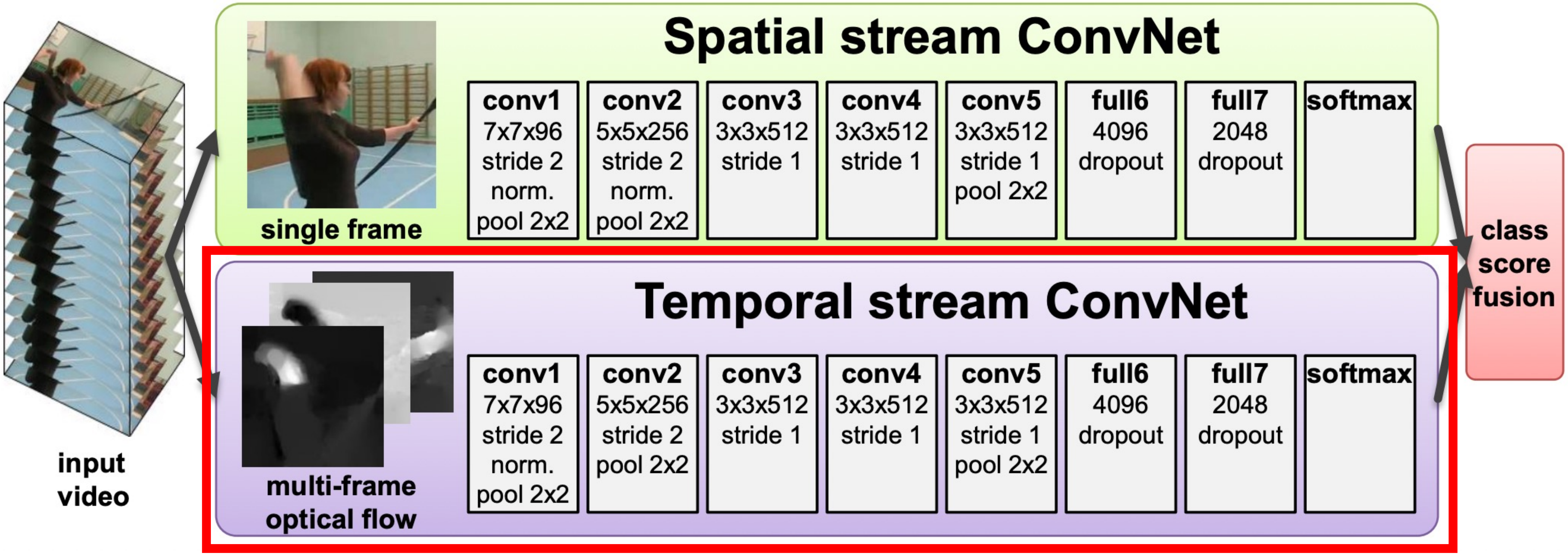
Two-Stream Architecture

Pre-trained on ImageNet



Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

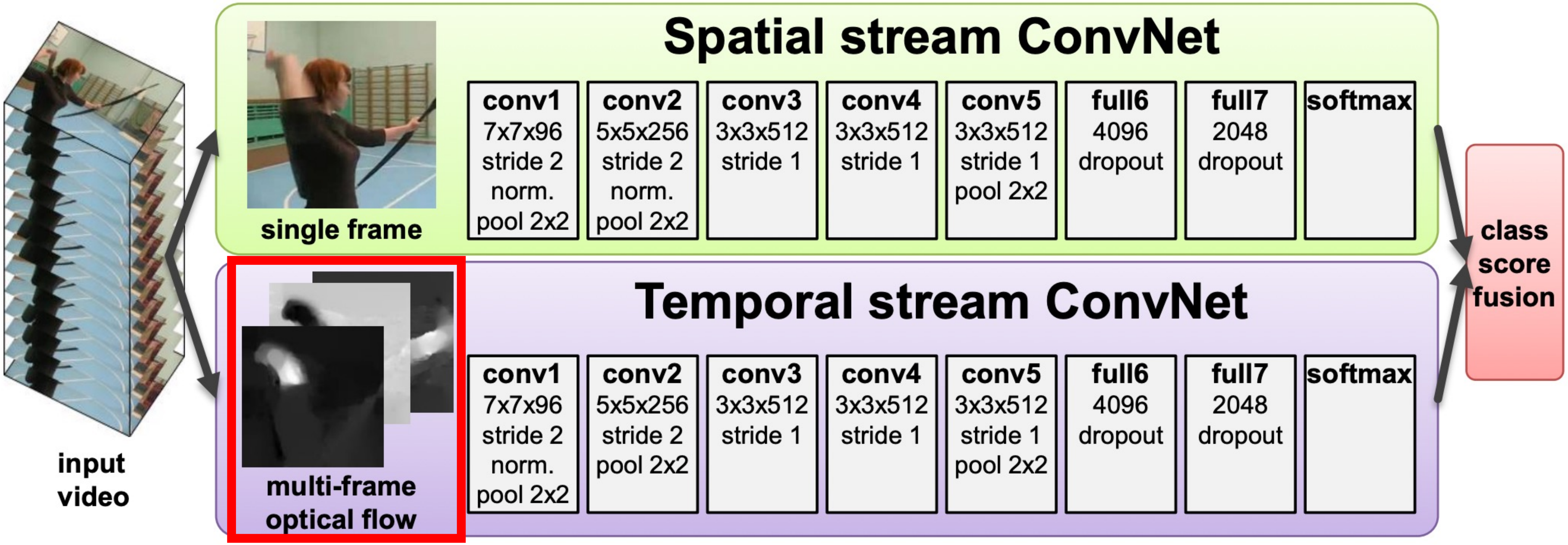
Two-Stream Architecture



Learns to predict from explicit motion representations the actions

Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture

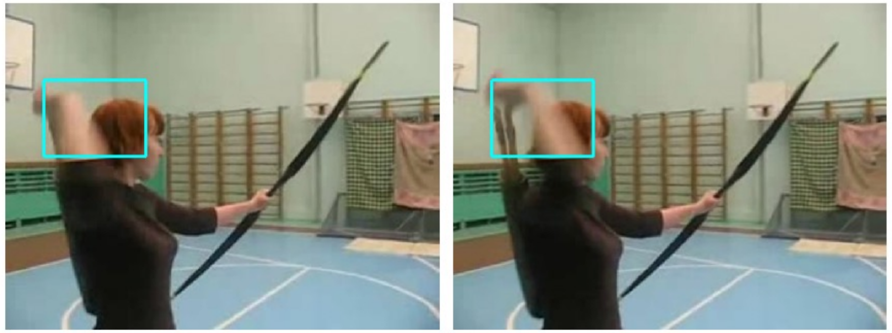


Input

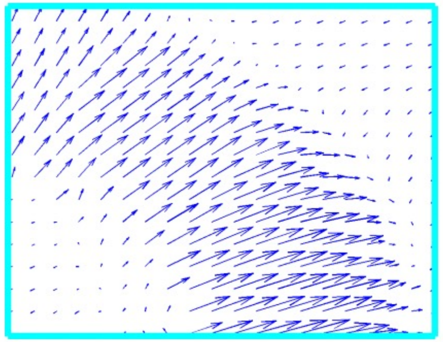
Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture: Input (Optical Flow)

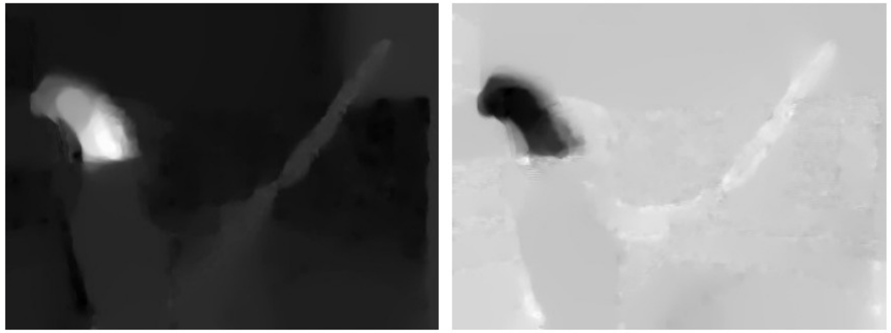
Input: stack pairs from consecutive frames



Two consecutive frames



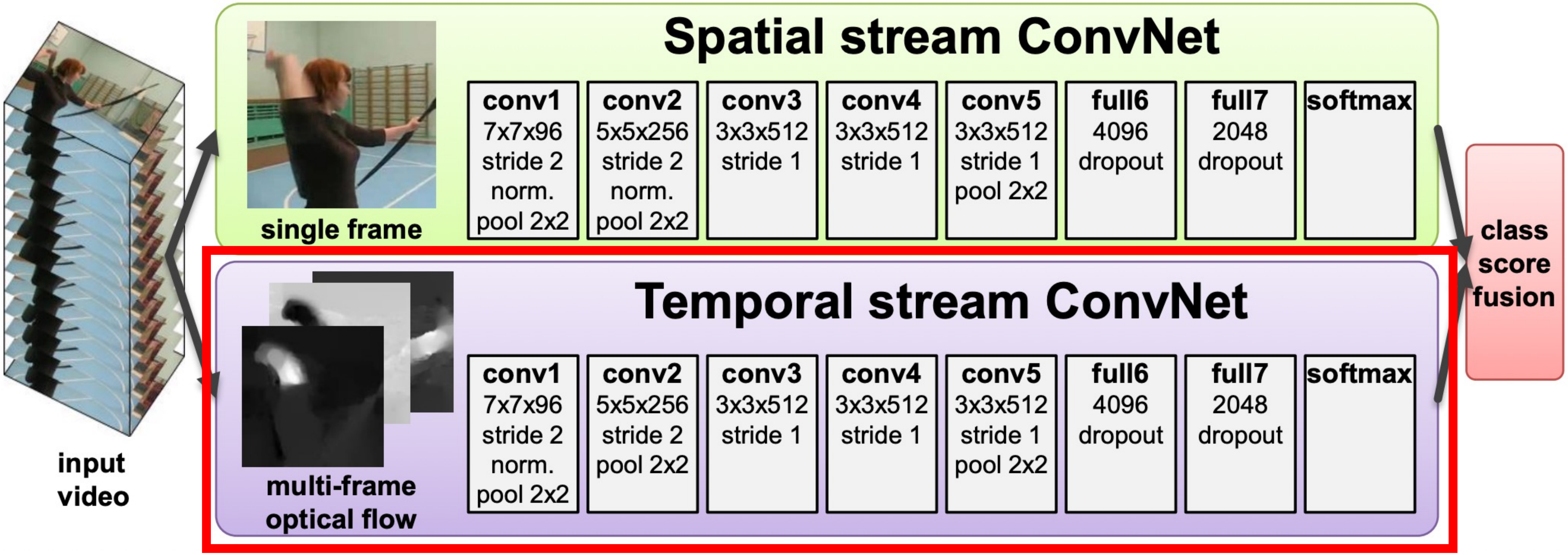
Vector fields showing where each point in the original frame moved in the subsequent frame



Vector fields decomposed into their horizontal (left) and vertical components (right)

Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture

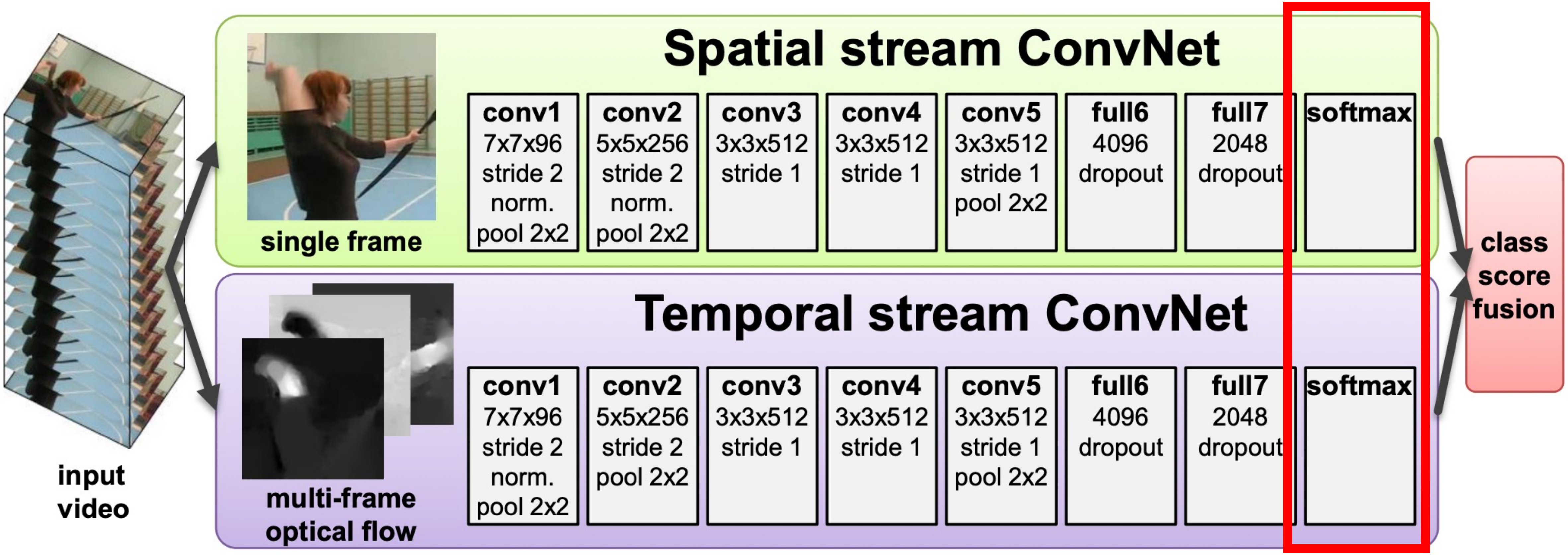


Must be trained on video datasets

Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture

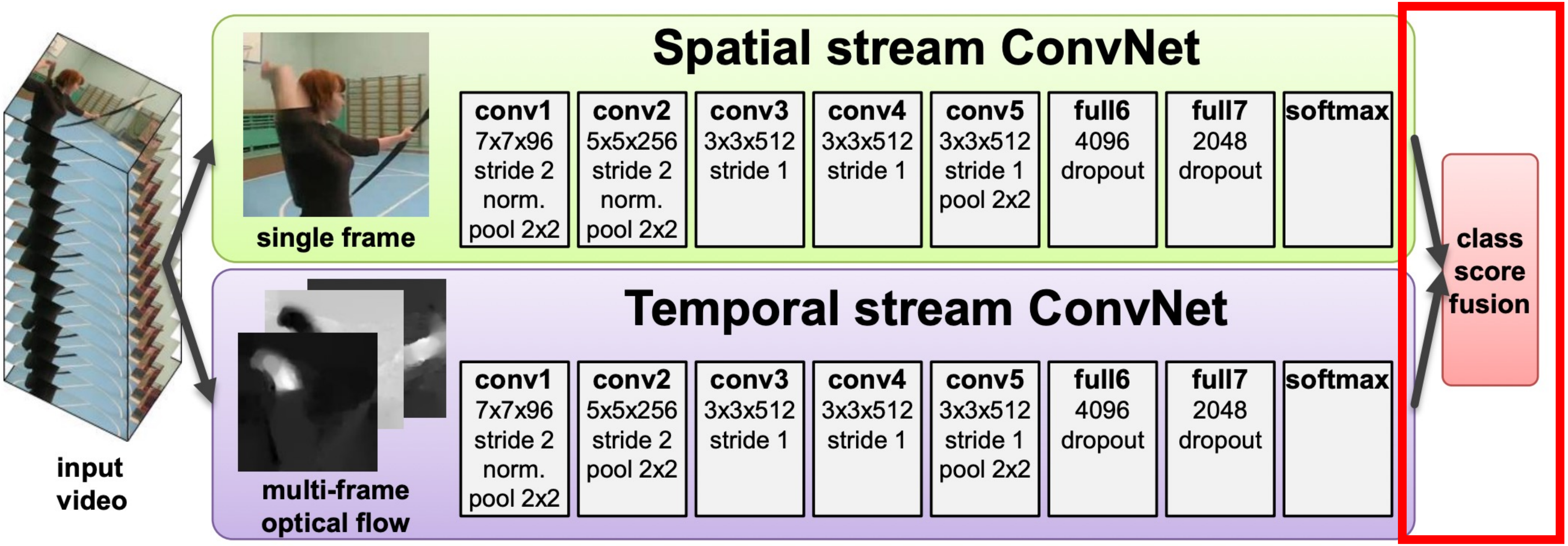
Both ConvNets learn to output a class score



Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture

Both class scores can be fused (e.g., averaging or using an SVM)

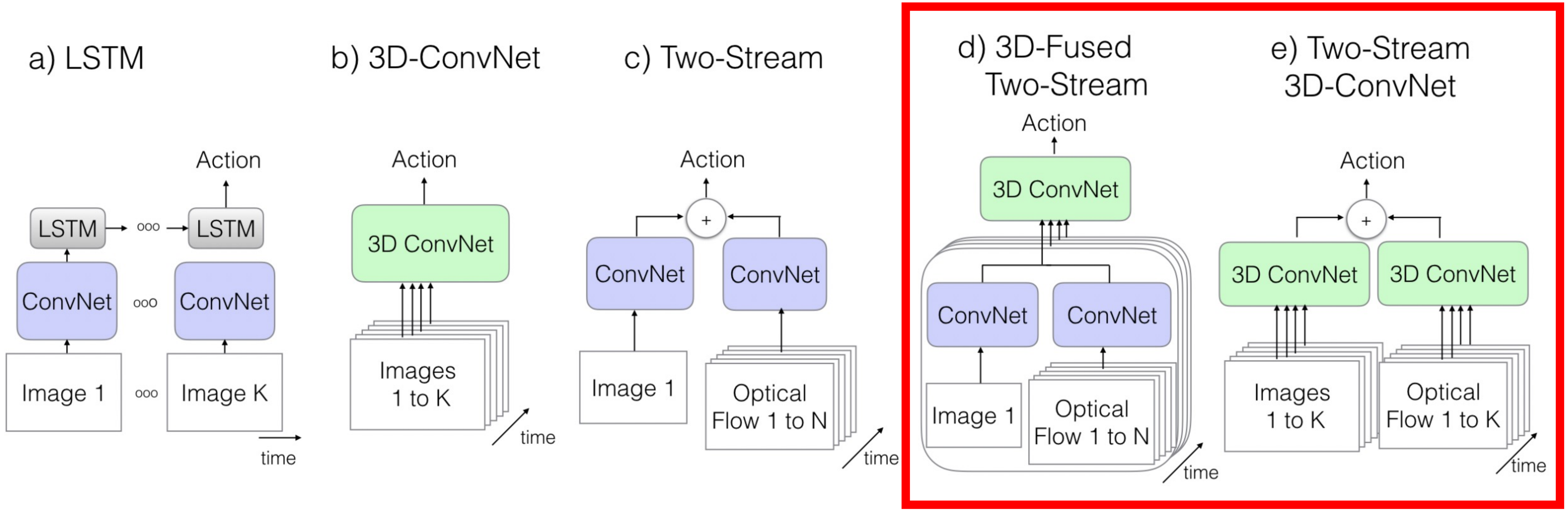


Karen Simonyan and Andrew Zisserman. Two-stream Convolutional Networks for Action Recognition in Videos. Neurips 2014.

Two-Stream Architecture: Limitations

- Successful training requires many videos which the architecture resource-hungry and time-consuming
- Does not capture long-term temporal information (duration determined by sequence duration used for the temporal stream ConvNet)
- Motion representation is limited by the assumptions of optical flow (e.g., constant appearance and smooth flow between frames)

Approaches to Capture Temporal Information: Can Also Mix Basic Approaches



Source: Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017.

Video Classification: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models


The End