

Video Understanding using CNNs and RNNs

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad



Review: Questions

How does GRU address vanishing gradients?

Same reason as the LSTM. There is a gradient highway, affected only by the update gate (which controls gradients by design and necessity)

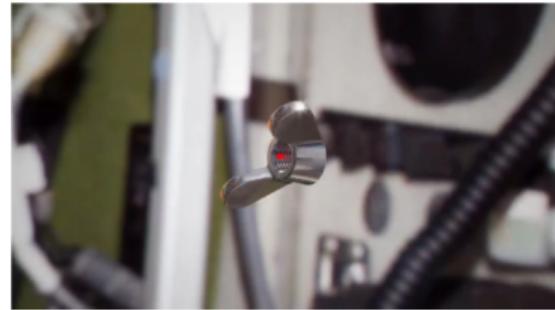
NPTEL

Why do we need to understand a video?



Credit: Smarter Everyday (Youtube)

Why do we need to understand a video?



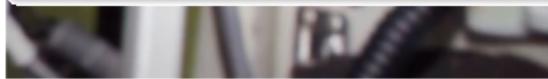
Credit: Veritasium (Youtube)

Why do we need to understand a video?



How to understand a video?

Let's forget everything we learn't and see if we can figure it out by ourself!



NPTEL

How to understand a video?



How to understand a video?

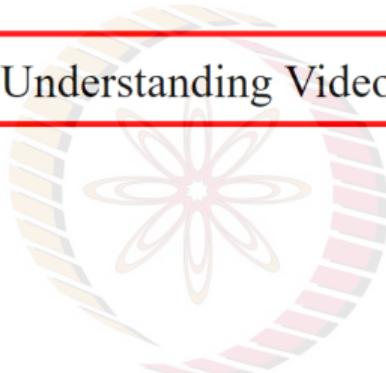


How to understand a video?



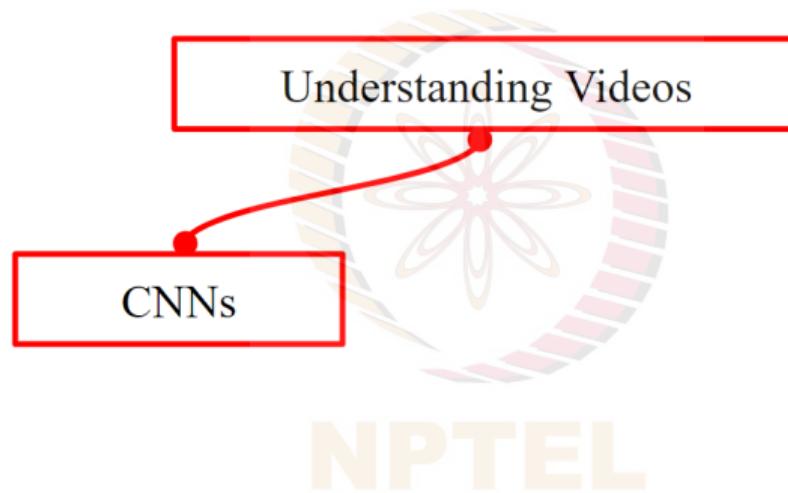
How to understand a video?

Understanding Videos

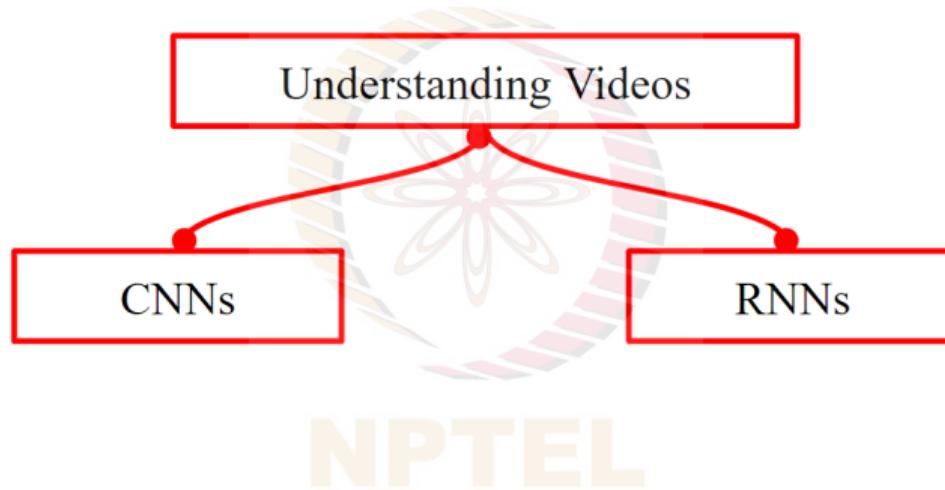


NPTEL

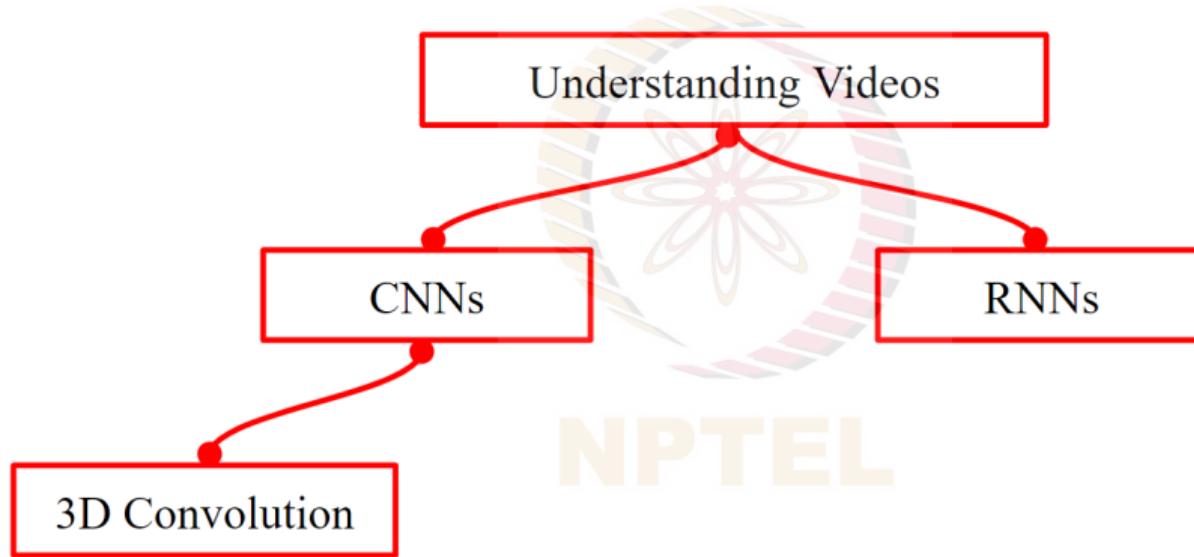
How to understand a video?



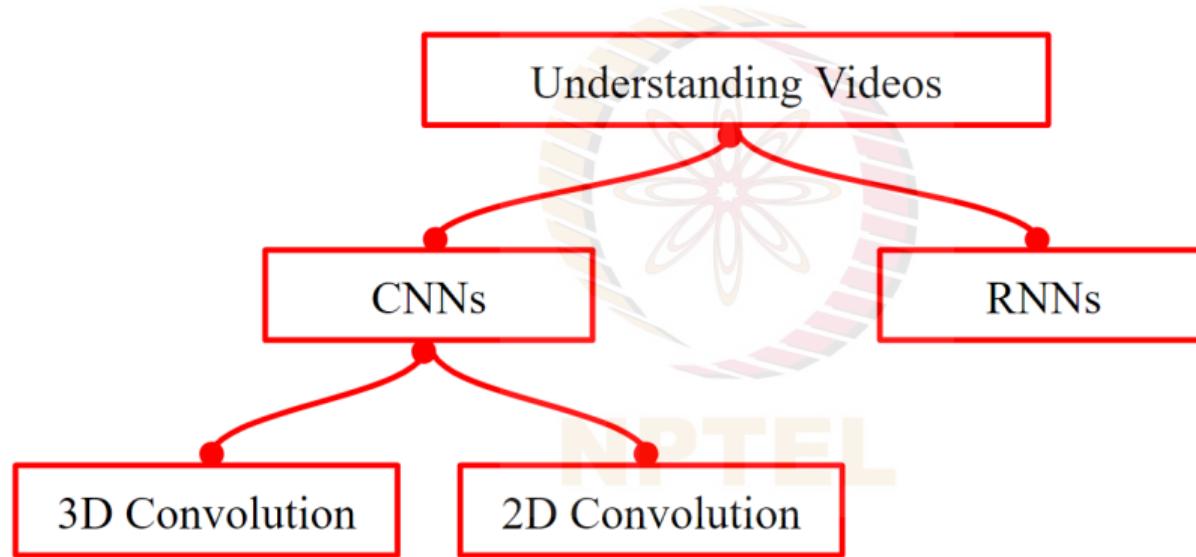
How to understand a video?



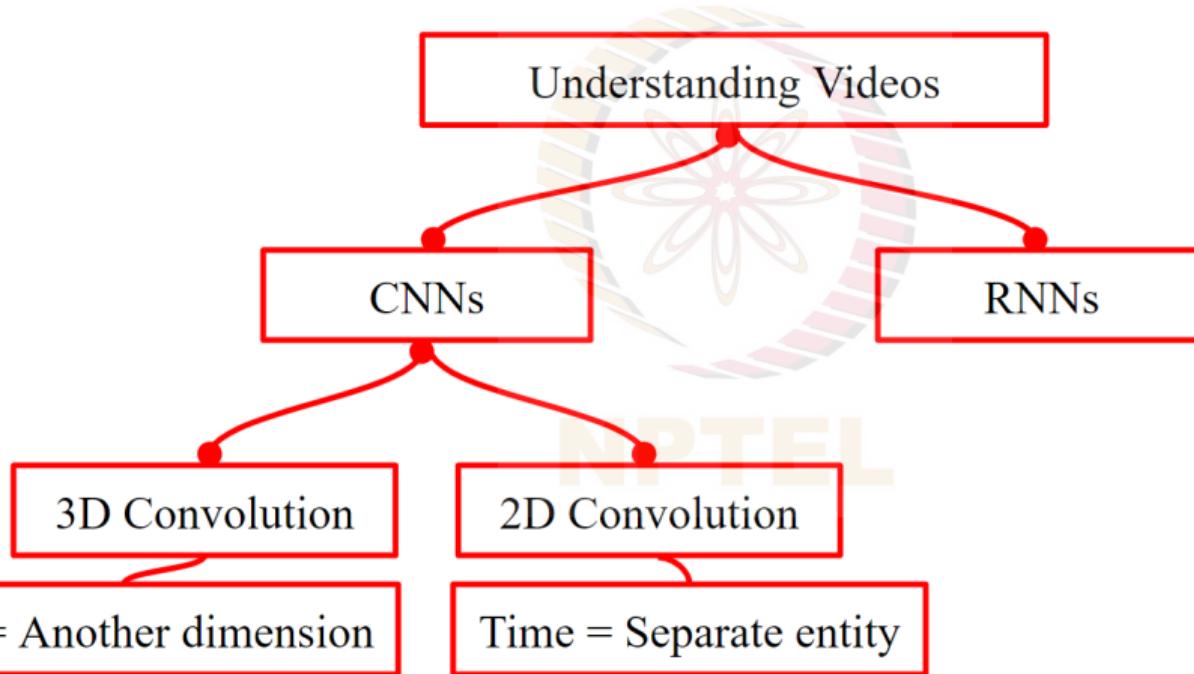
How to understand a video?



How to understand a video?



How to understand a video?



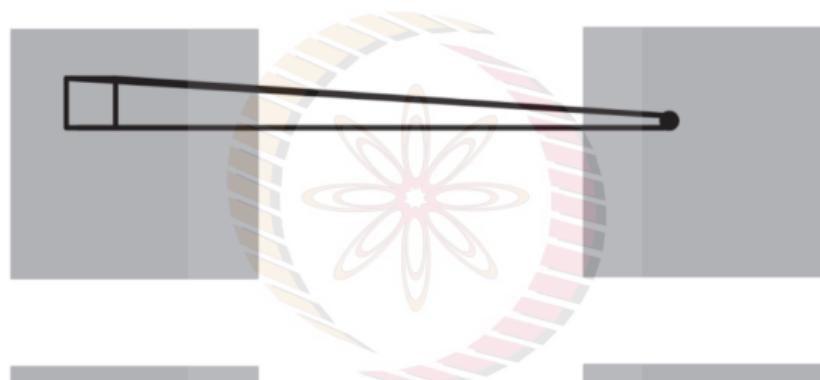
How to understand a video? 3D CNN

Frame 1

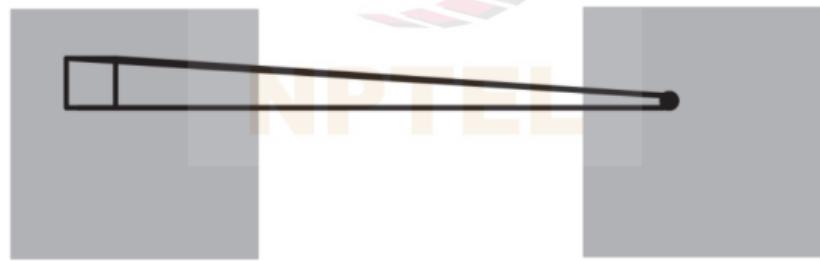


How to understand a video? 3D CNN

Frame 1



Frame 2

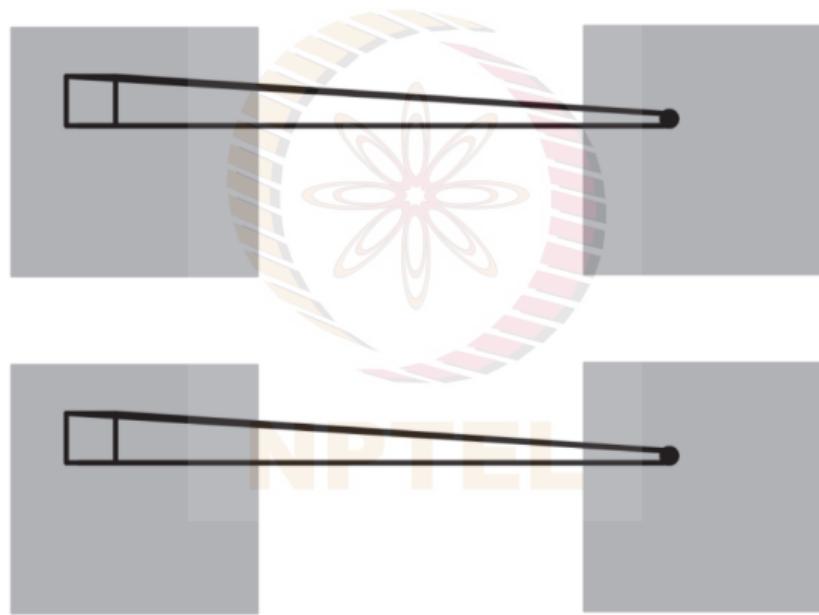


How to understand a video? 3D CNN

Frame 1

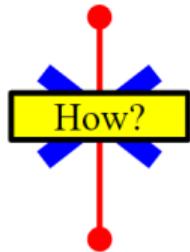


Frame 2

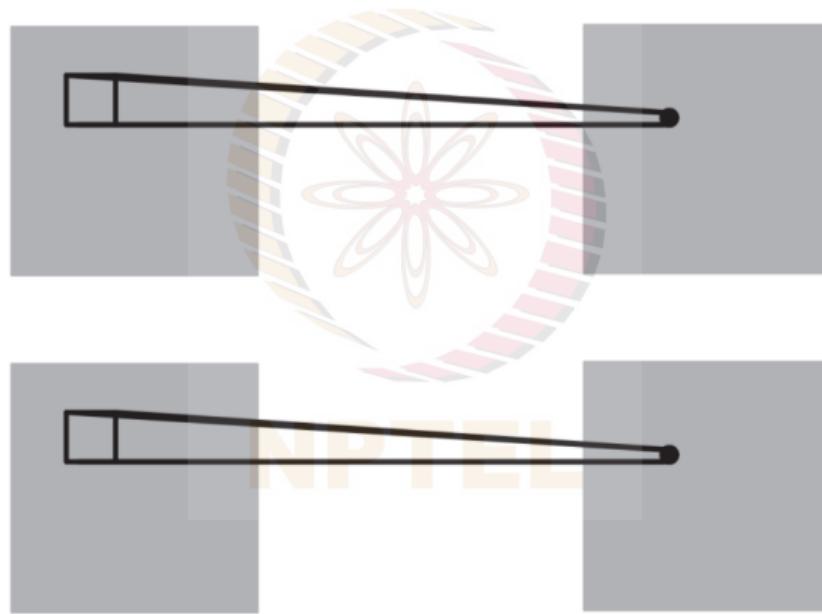


How to understand a video? 3D CNN

Frame 1

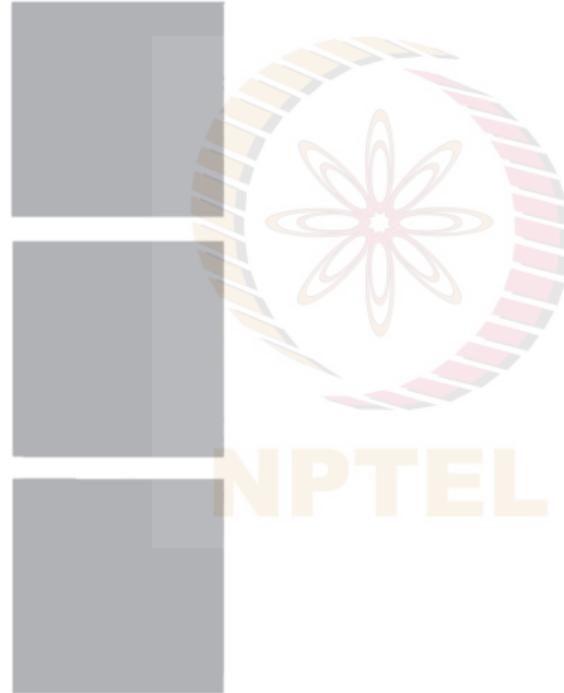


Frame 2



How to understand a video? 3D CNN

Frame 1



Frame 2

Frame 3

How to understand a video? 3D CNN

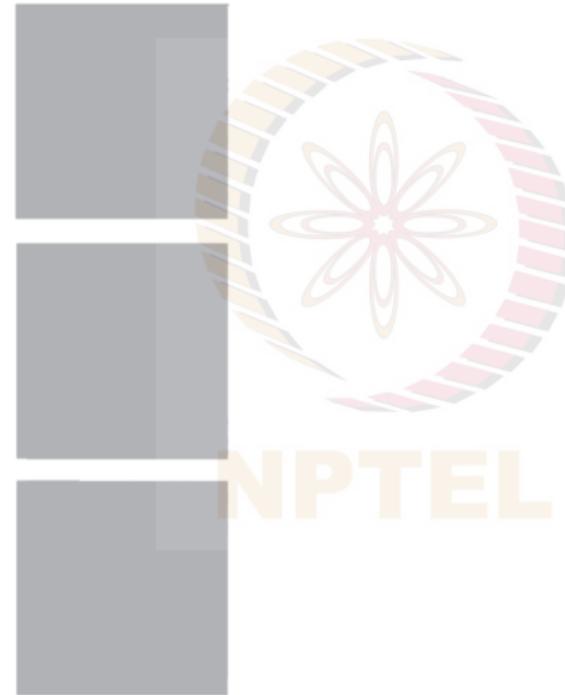
Frame 1



Frame 2



Frame 3



How to understand a video? 3D CNN

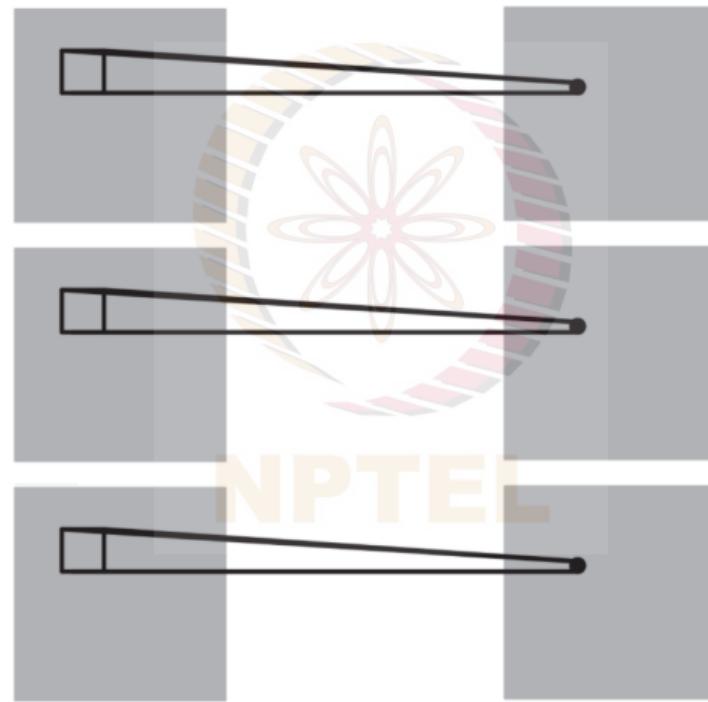
Frame 1



Frame 2



Frame 3



How to understand a video? 3D CNN¹

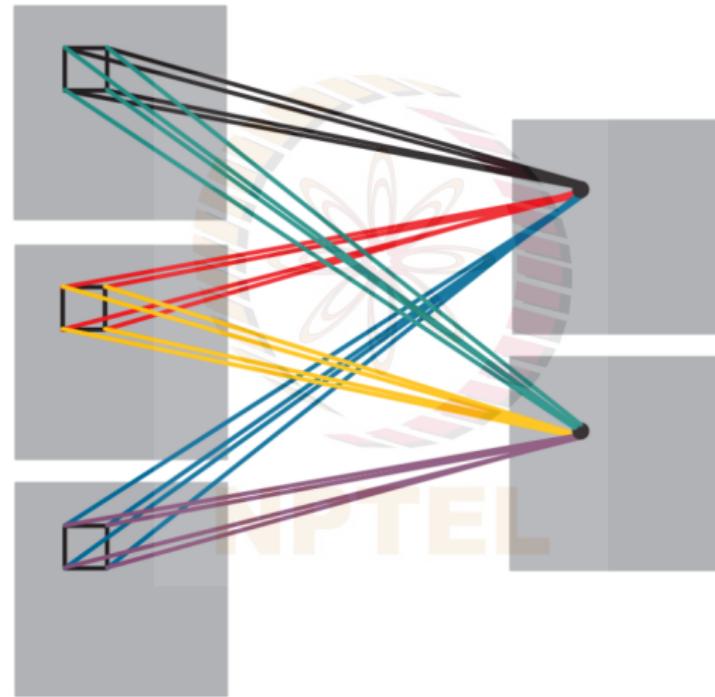
Frame 1



Frame 2

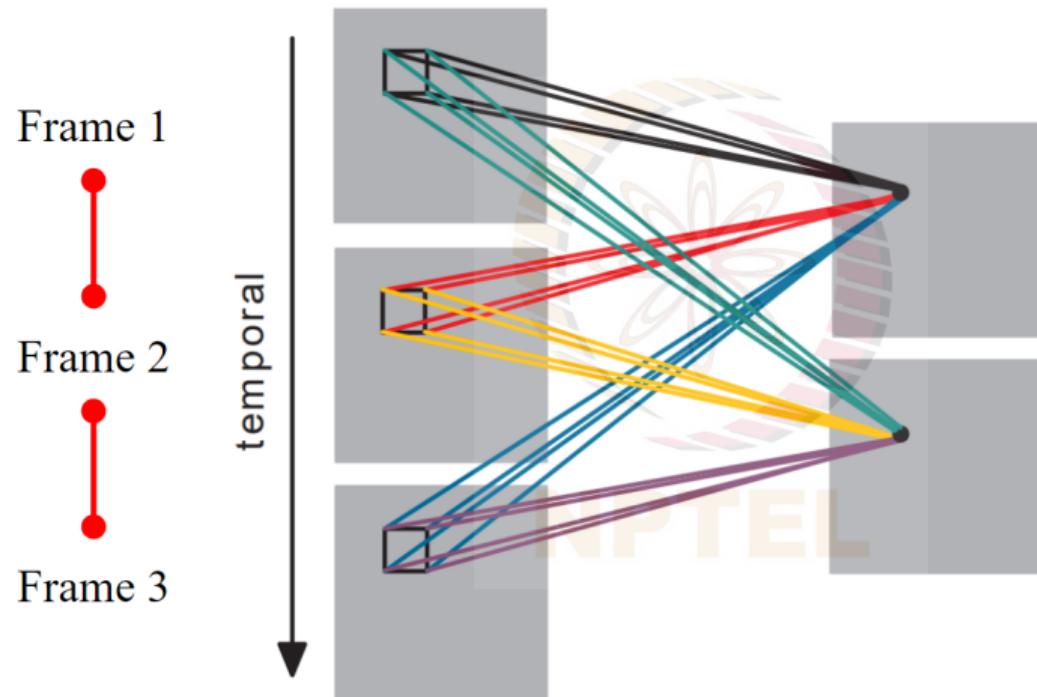


Frame 3



¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



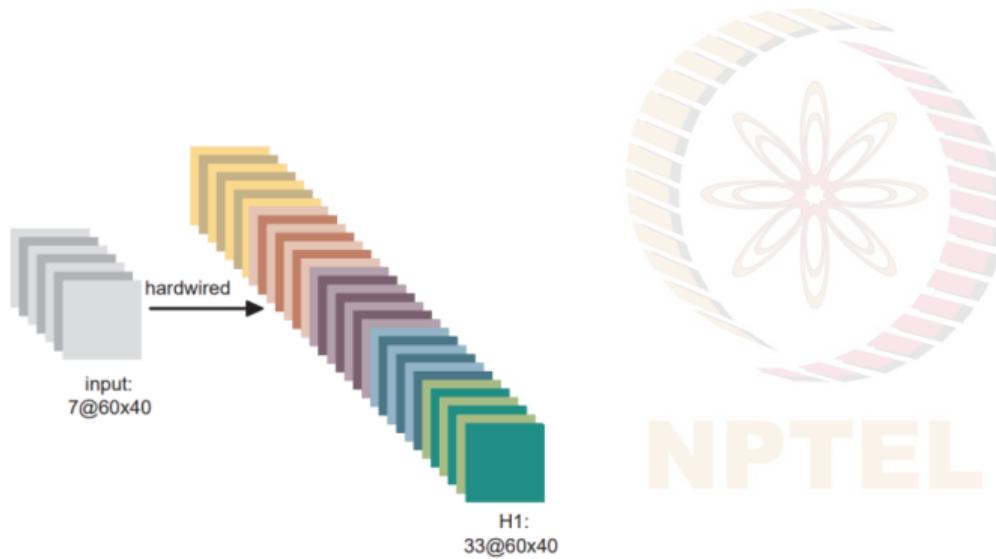
input:
7@60x40



NPTEL

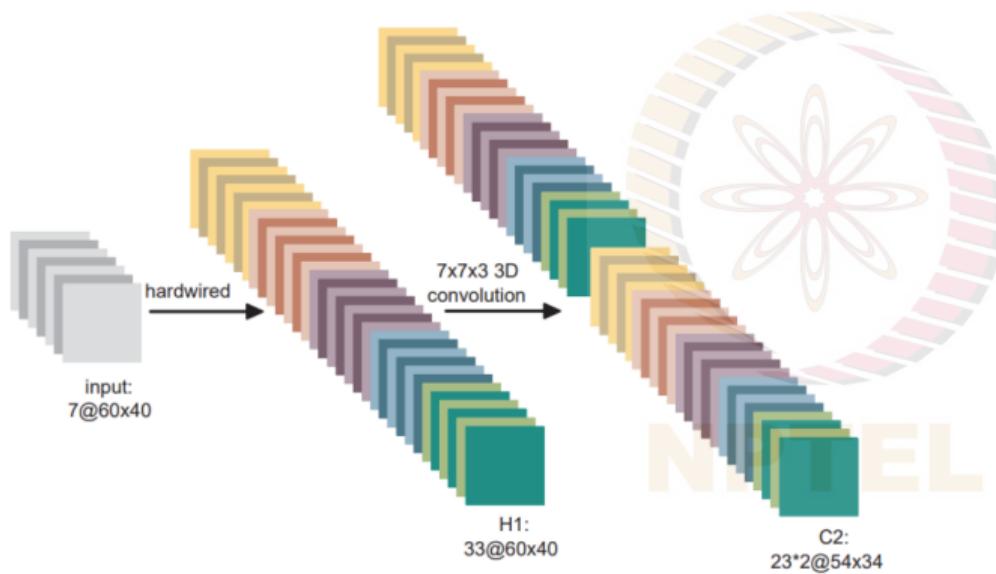
¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



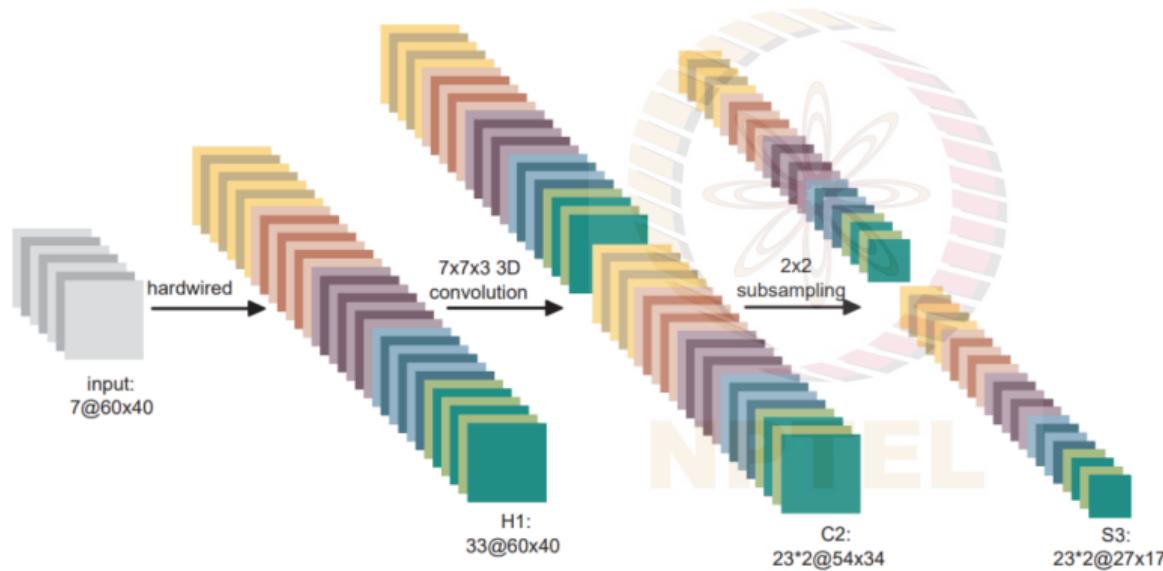
¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



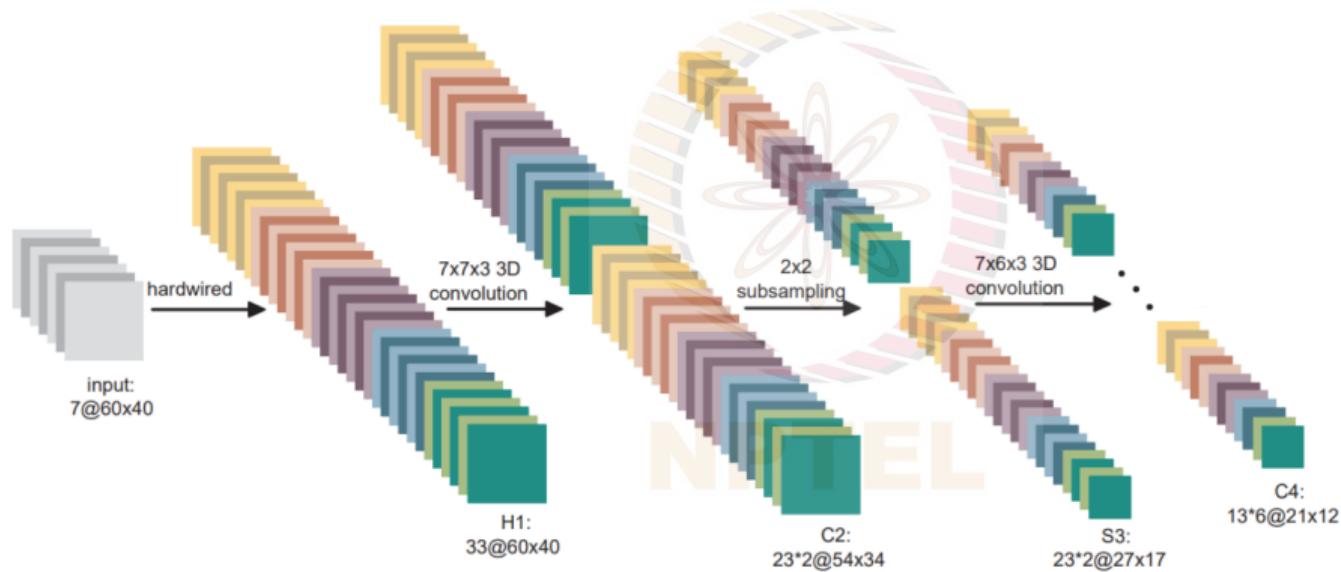
¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



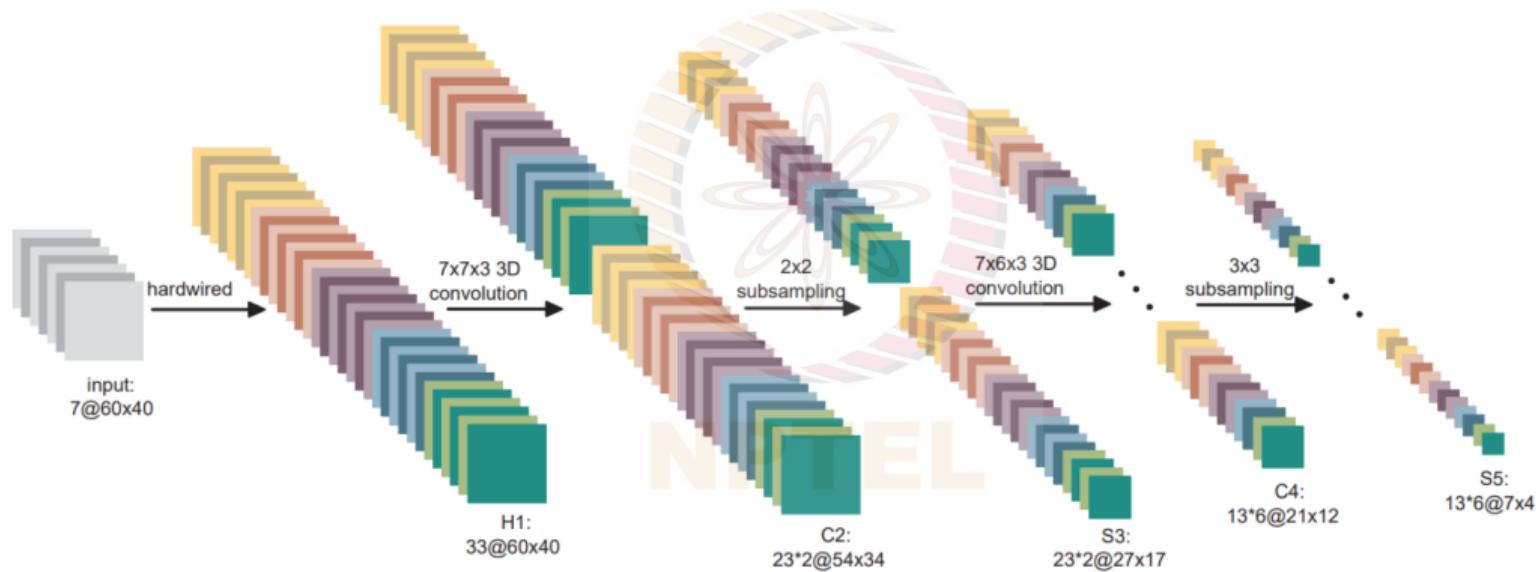
¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



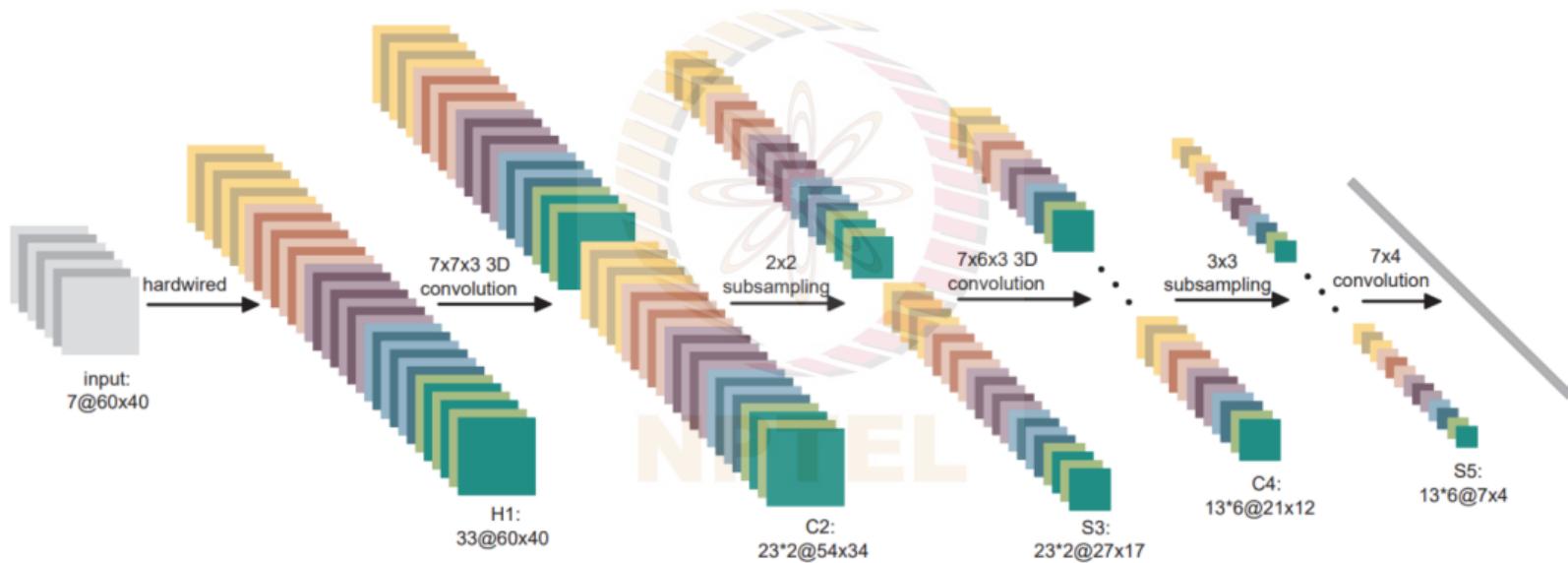
¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



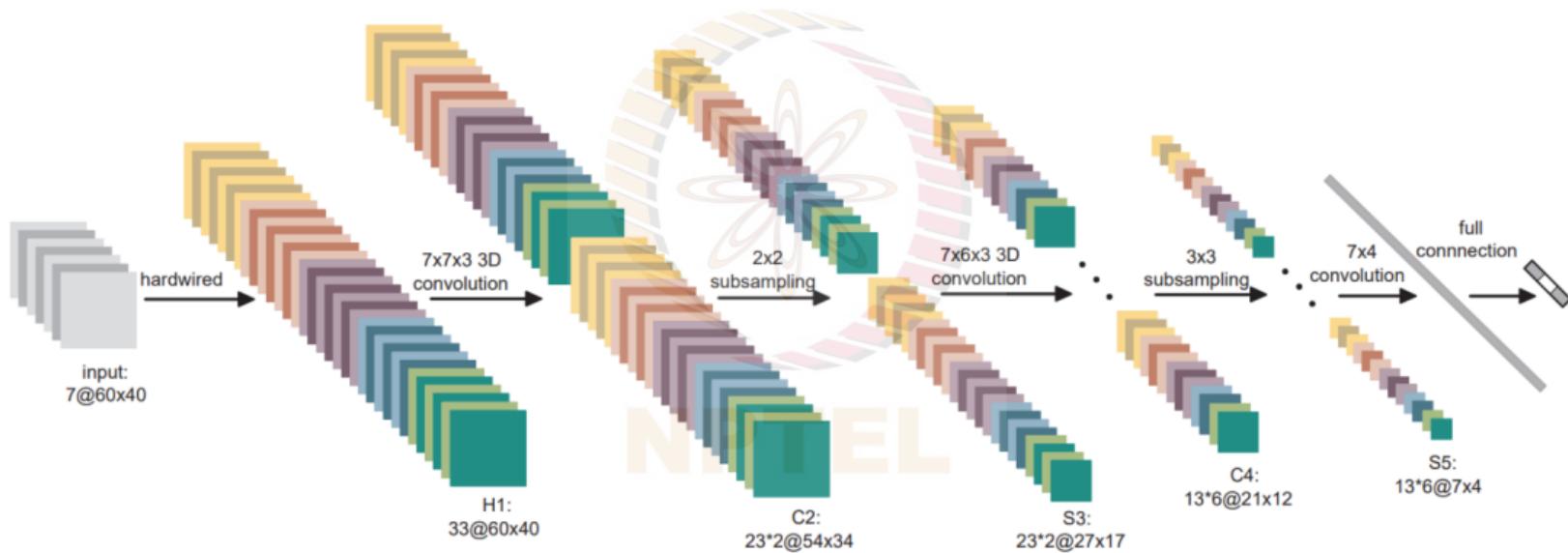
¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 3D CNN¹



CellToEar - Someone puts a cell phone to his/her head or ear.

ObjectPut - Someone drops or puts down an object.

Pointing - Someone points

¹Ji et al, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on PAMI, 2012

How to understand a video? 2D CNN²



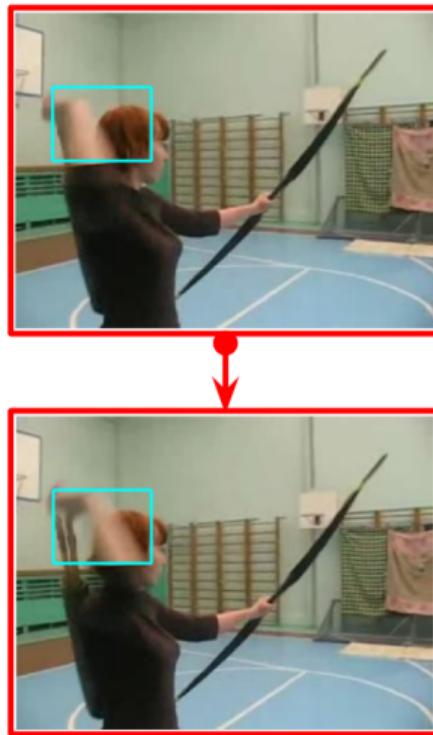
t_1



NPTEL

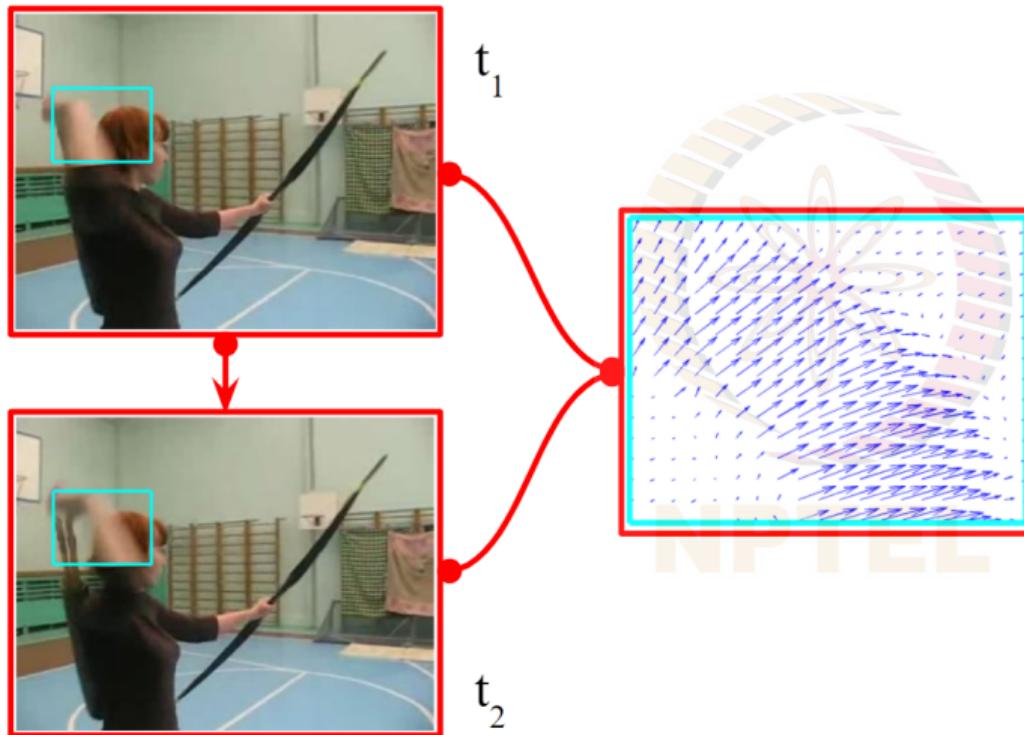
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



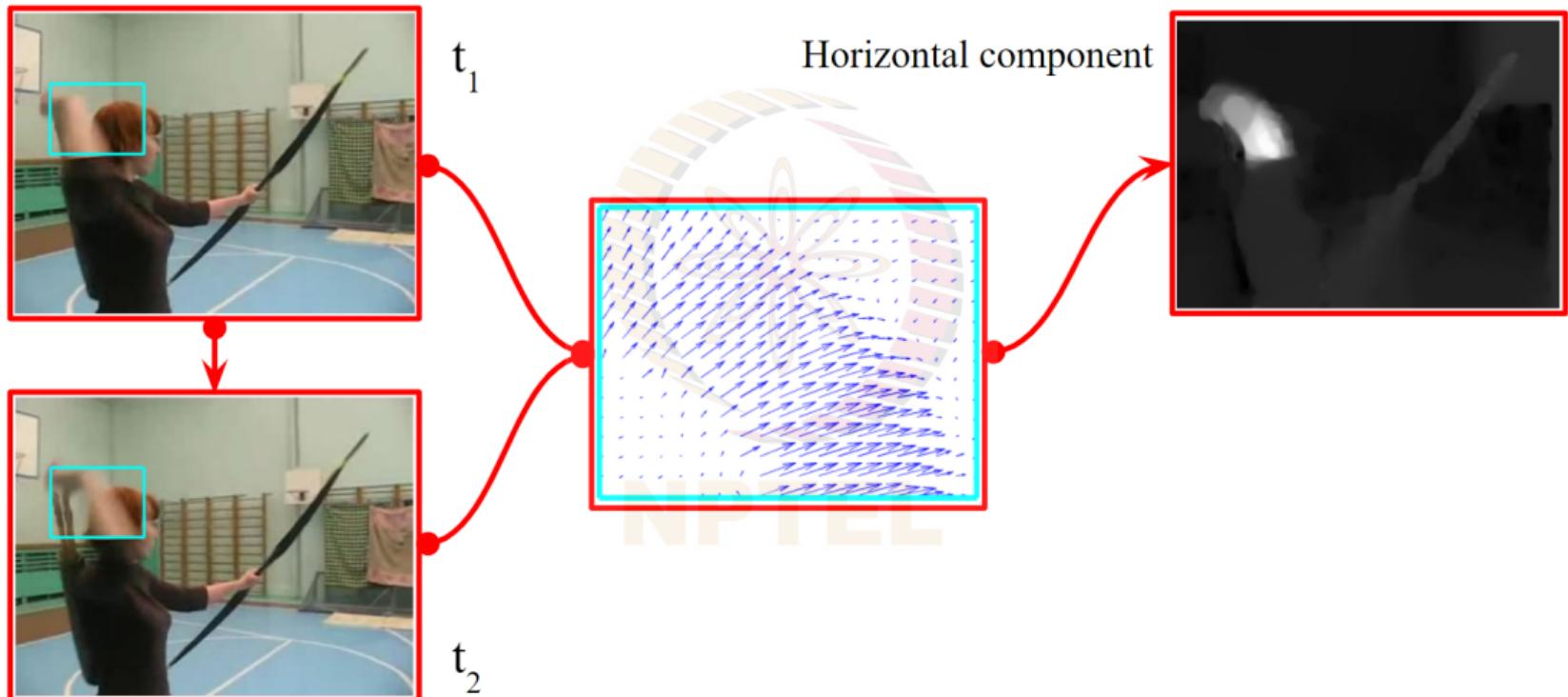
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



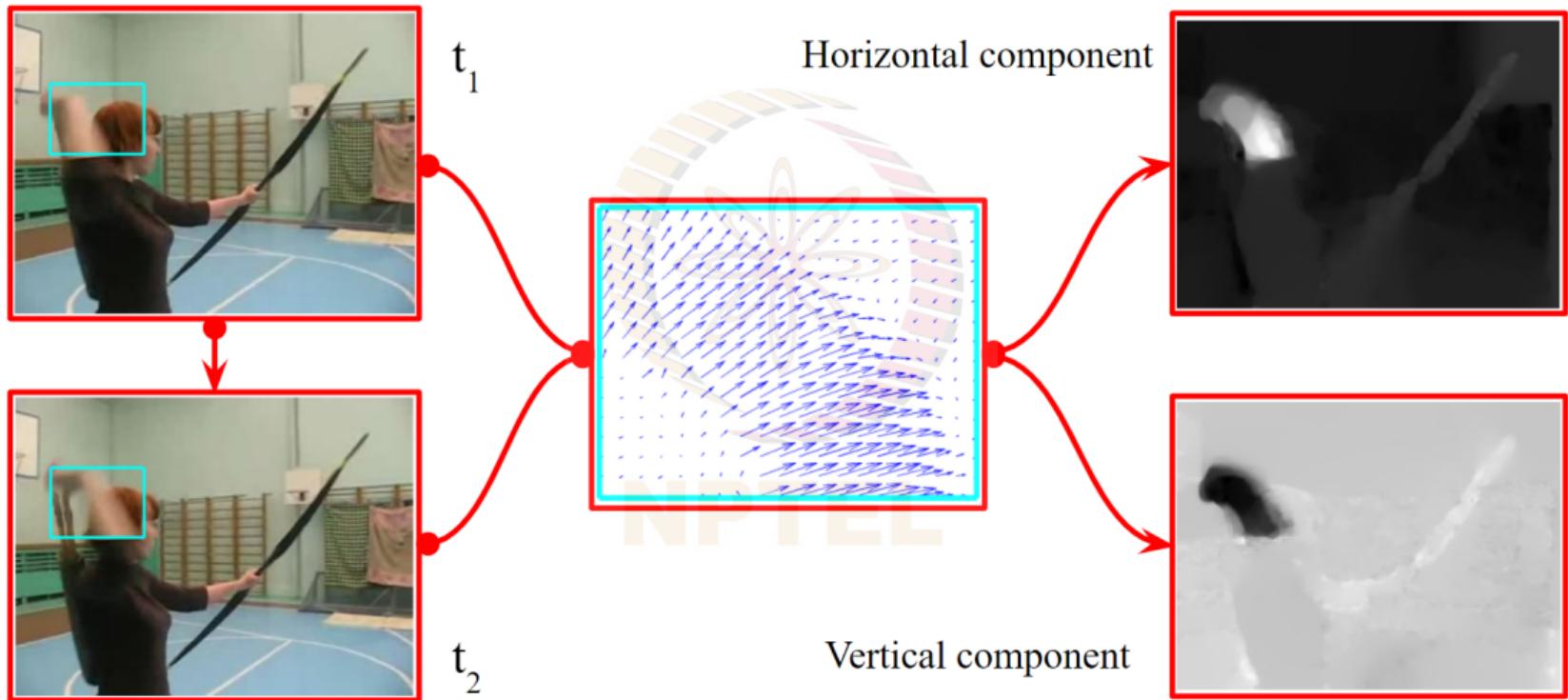
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



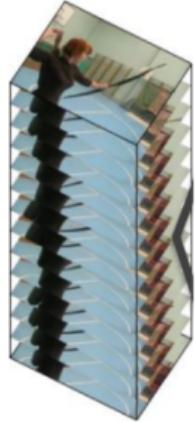
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²

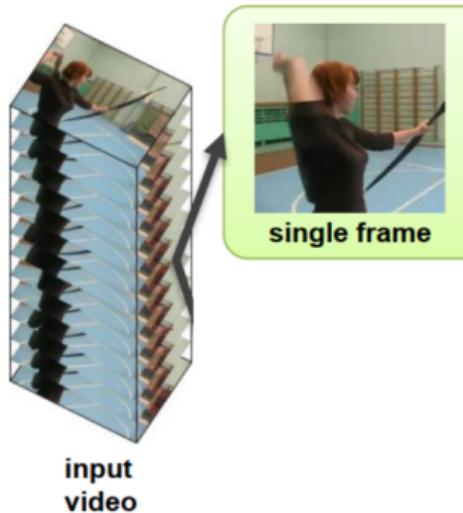


input
video



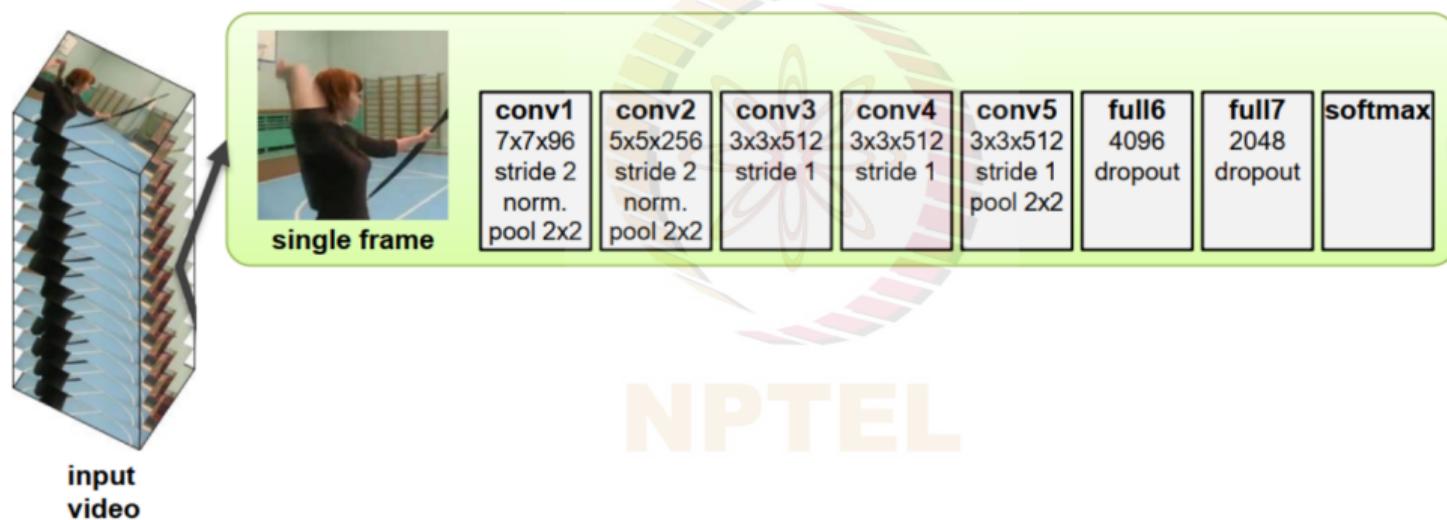
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



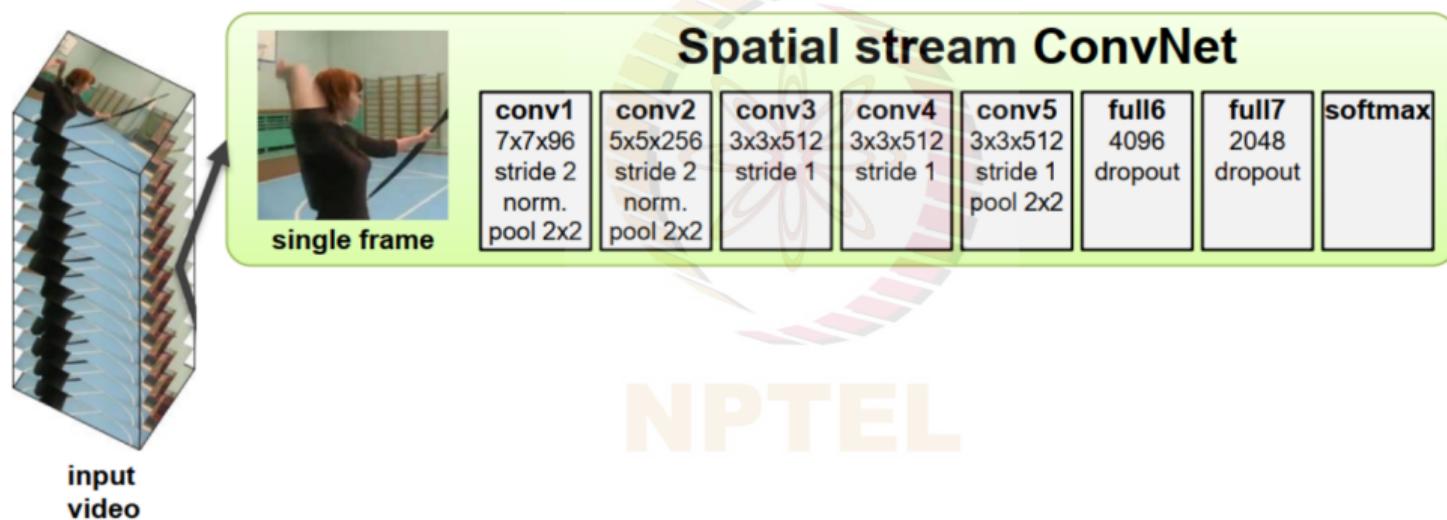
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

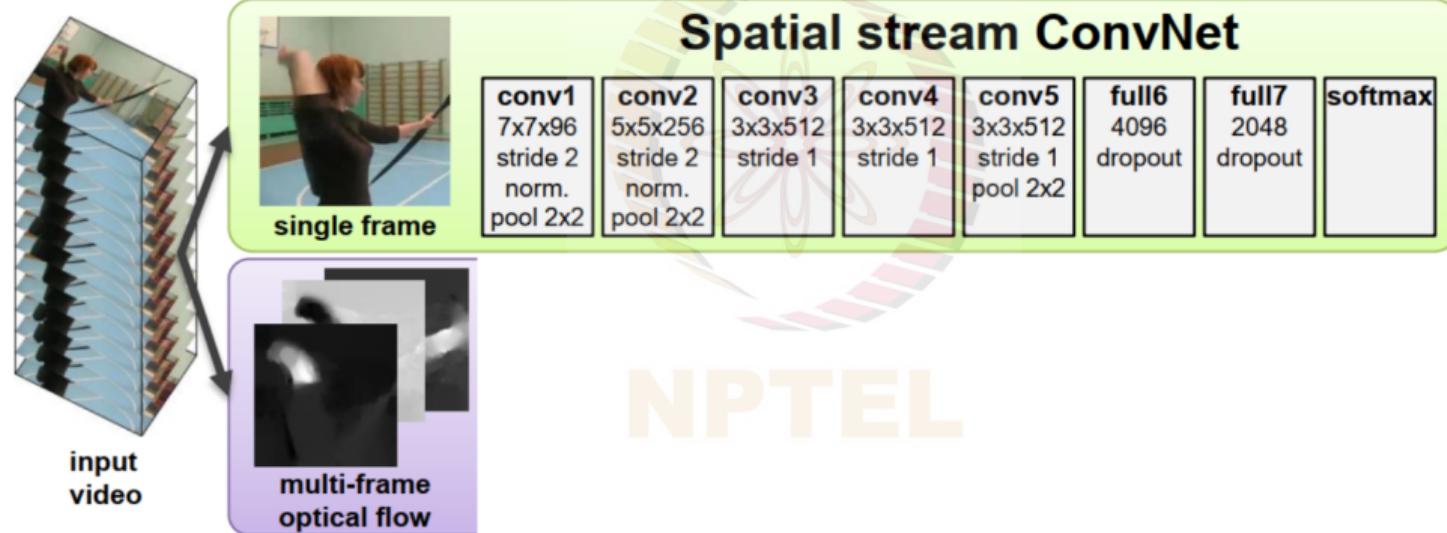
How to understand a video? 2D CNN²



NPTEL

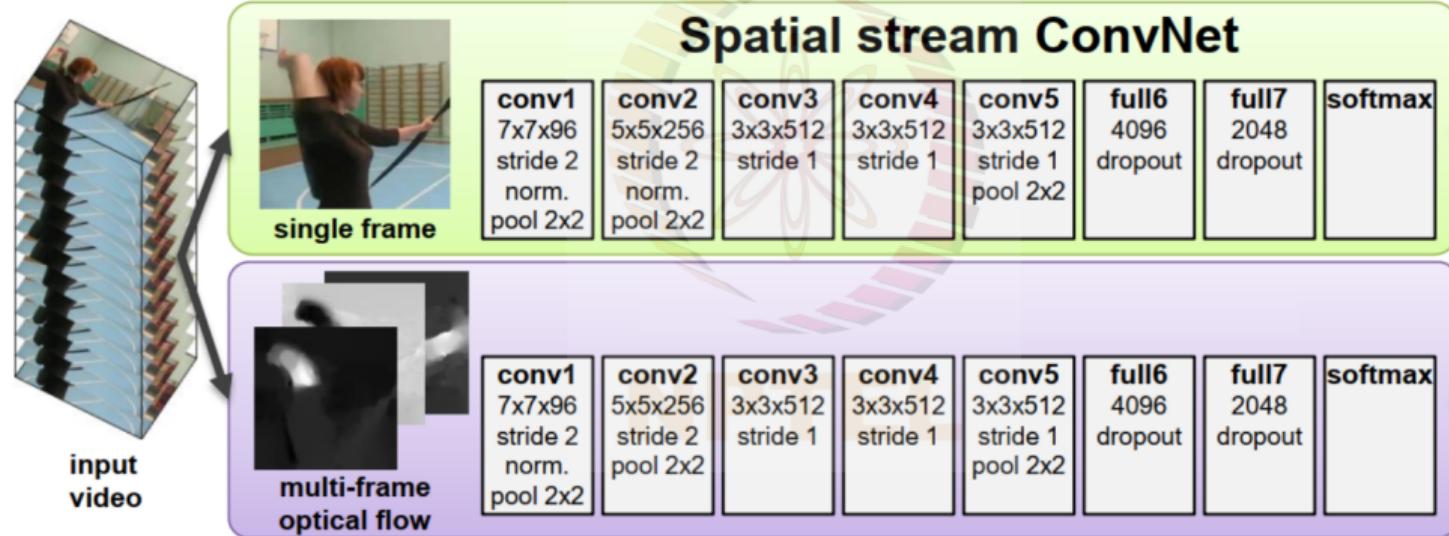
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



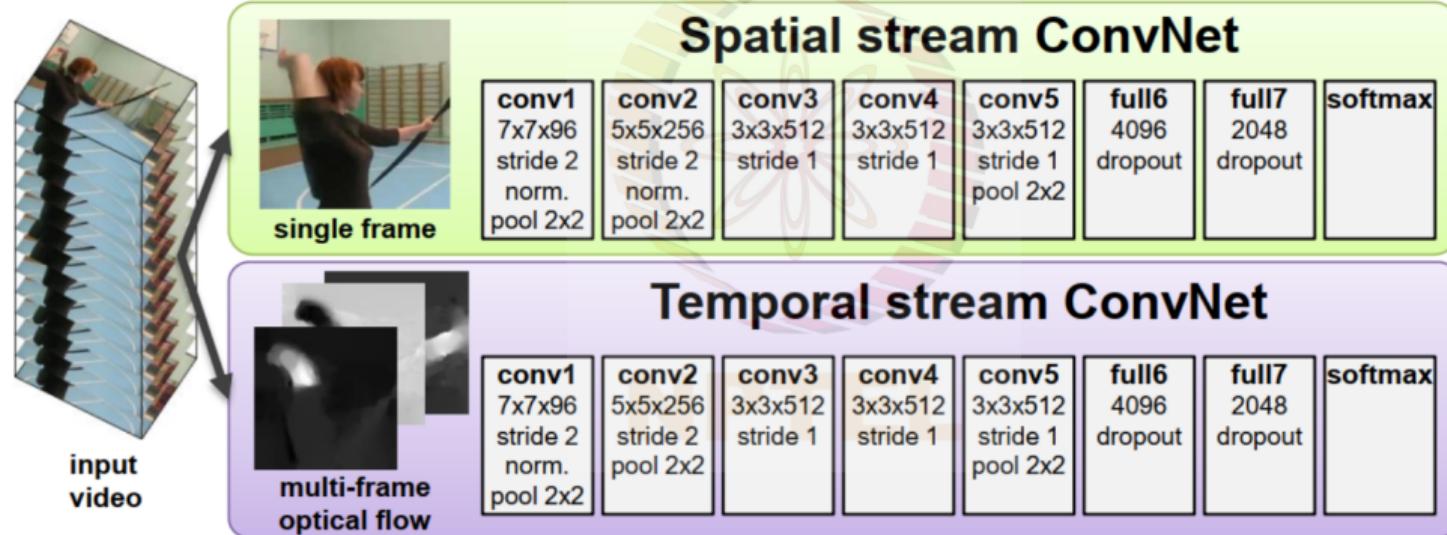
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



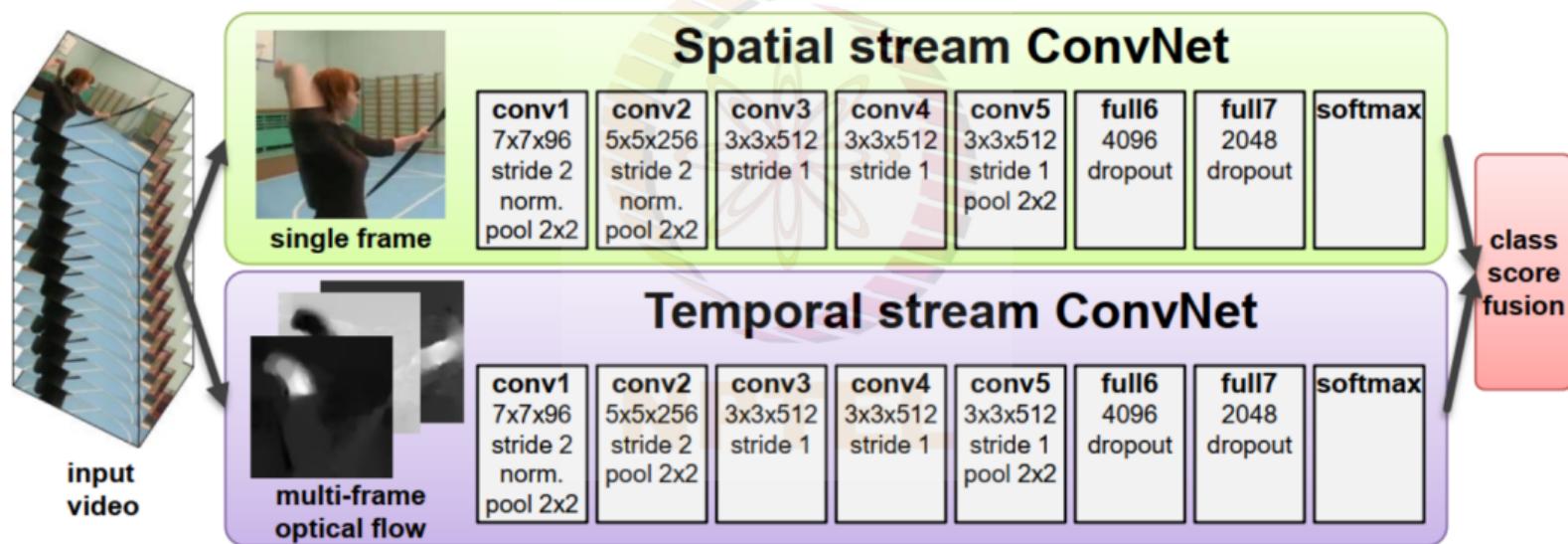
²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? 2D CNN²



²Simonyan and Zisserman, Two-stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

How to understand a video? Using RNNs with CNNs³

Visual Input

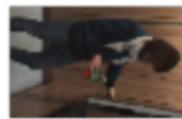


NPTEL

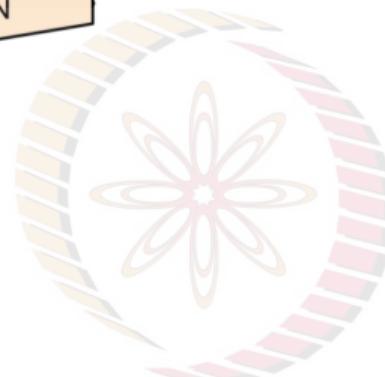
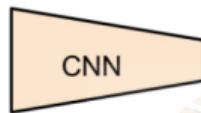
³Donahue et al, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

How to understand a video? Using RNNs with CNNs³

Visual Input



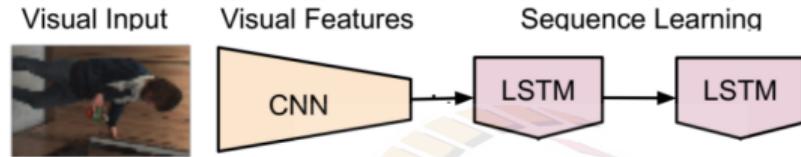
Visual Features



NPTEL

³Donahue et al, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

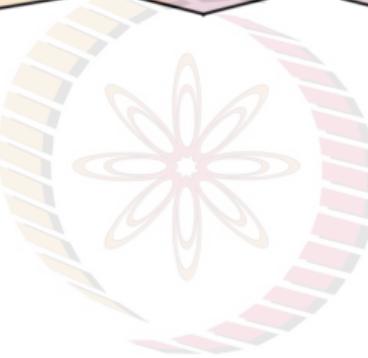
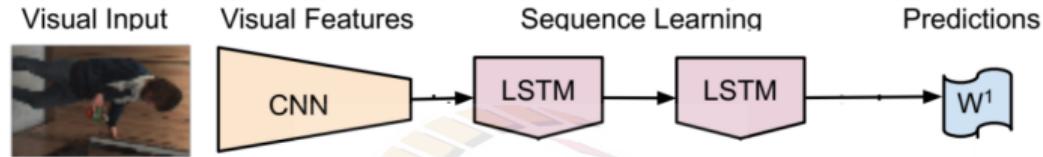
How to understand a video? Using RNNs with CNNs³



NPTEL

³Donahue et al, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

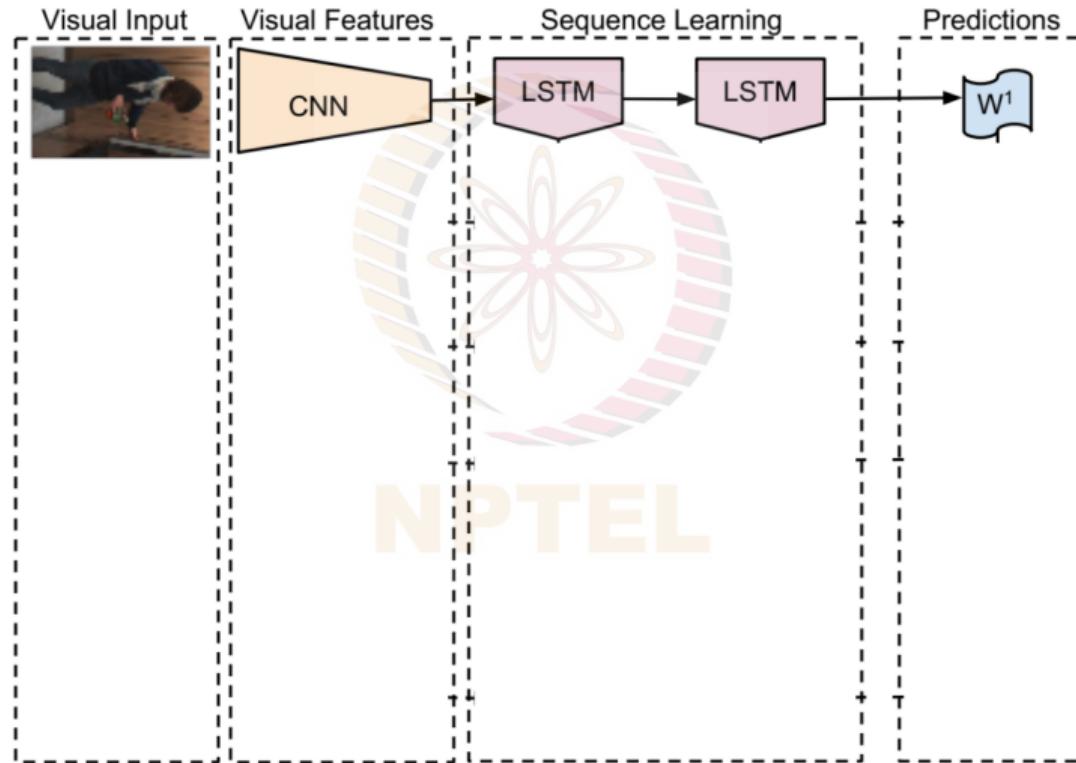
How to understand a video? Using RNNs with CNNs³



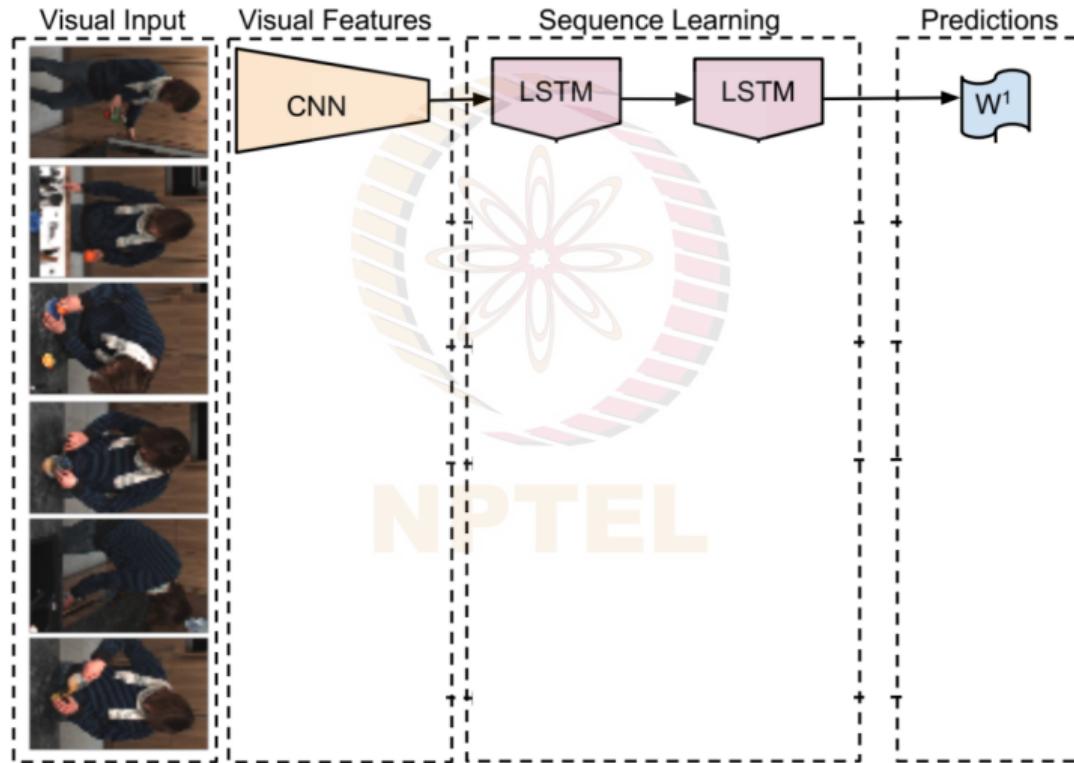
NPTEL

³Donahue et al, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

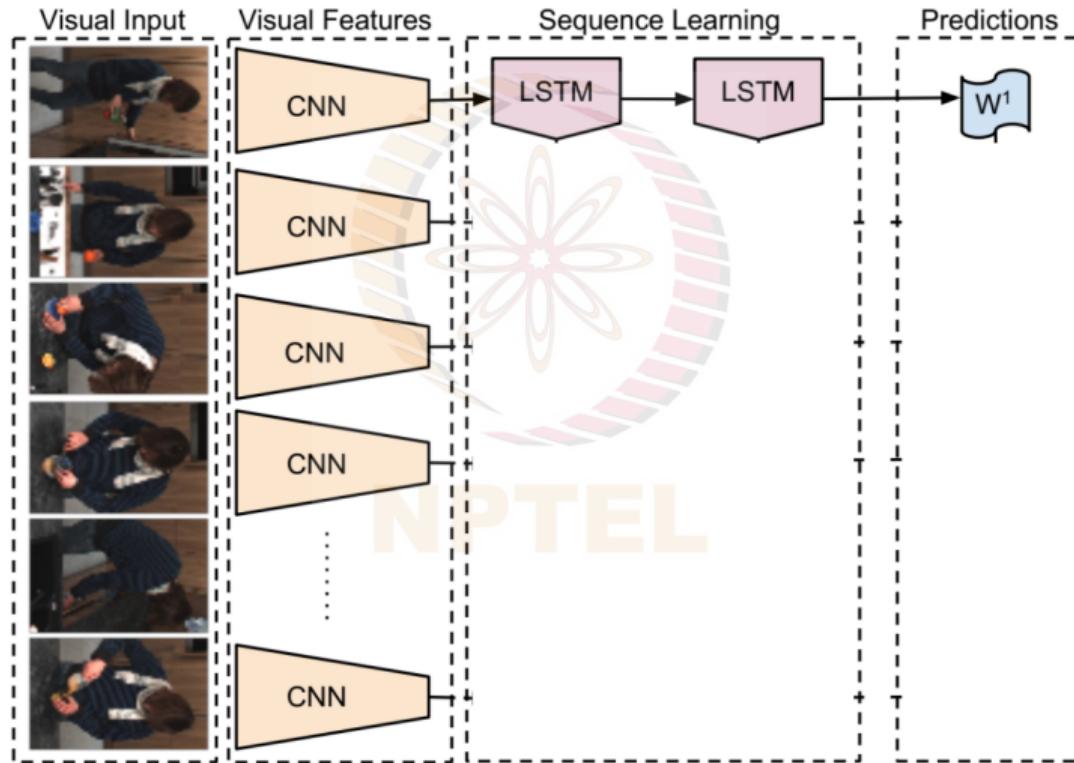
How to understand a video? Using RNNs with CNNs



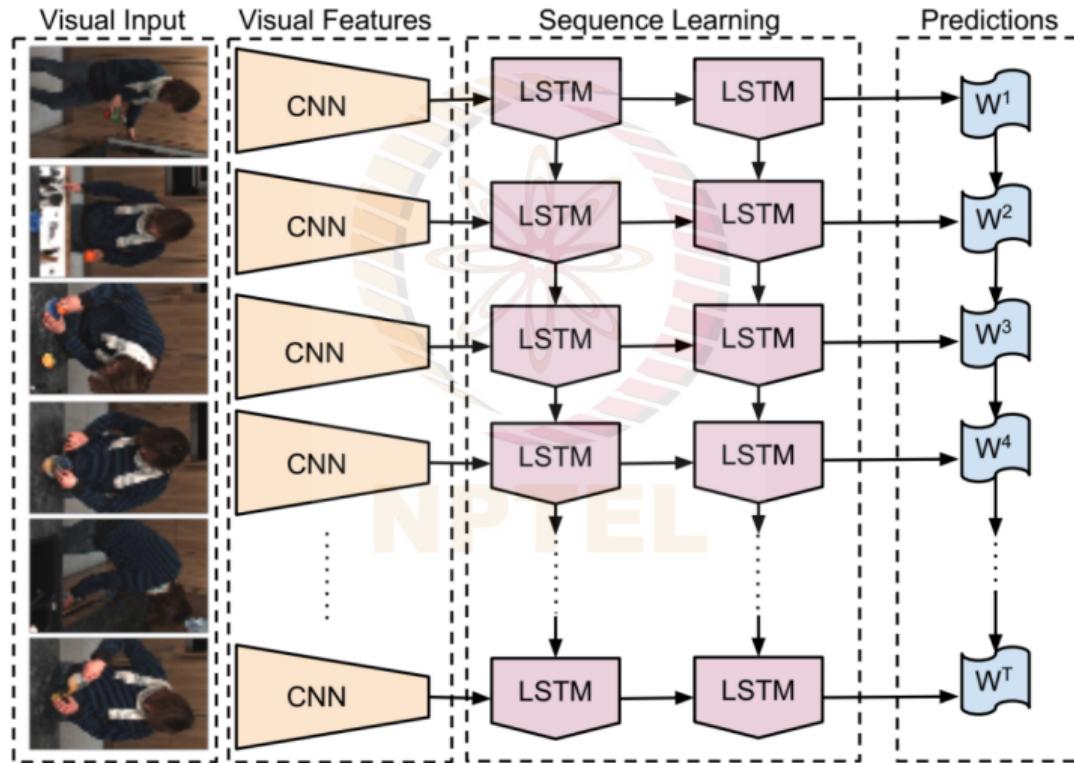
How to understand a video? Using RNNs with CNNs



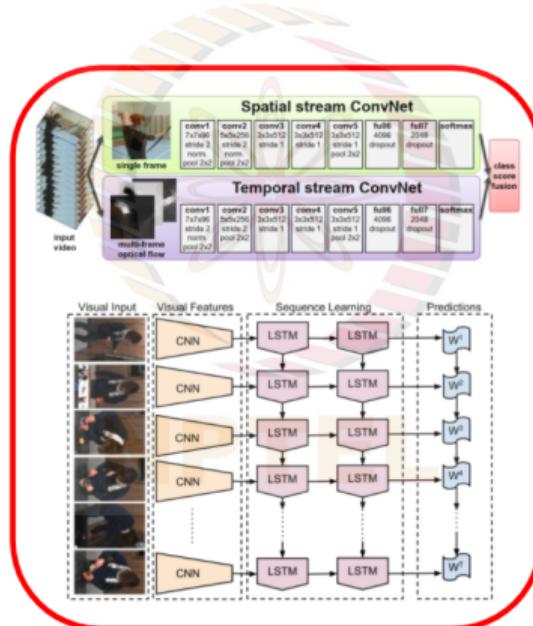
How to understand a video? Using RNNs with CNNs



How to understand a video? Using RNNs with CNNs



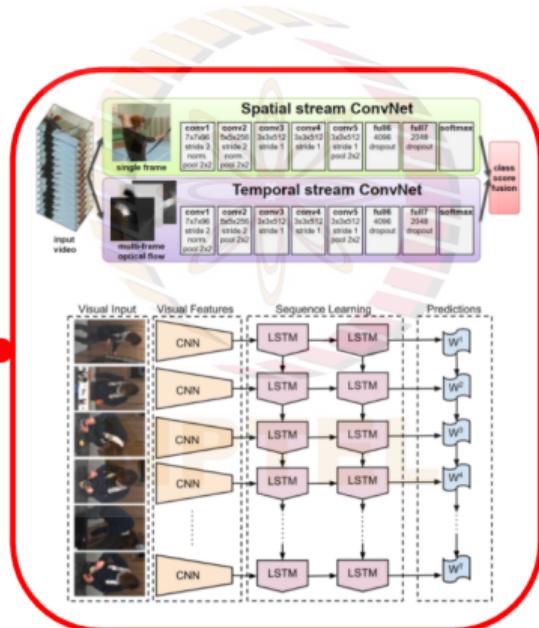
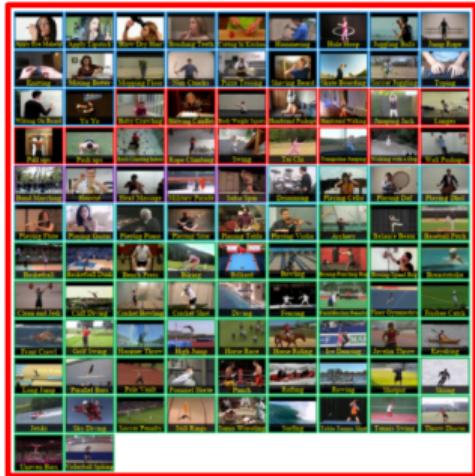
What can be done?



Soomro et al, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, 2012

What can be done?

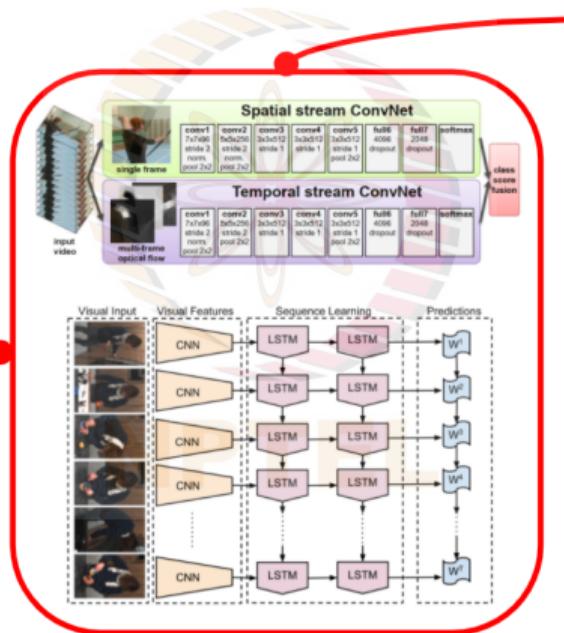
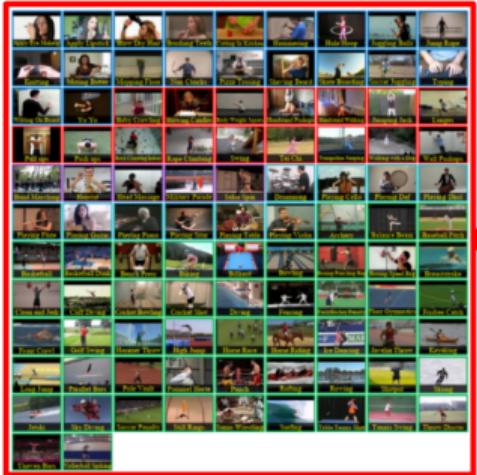
Train - UCF101



Soomro et al, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, 2012

What can be done?

Train - UCF101

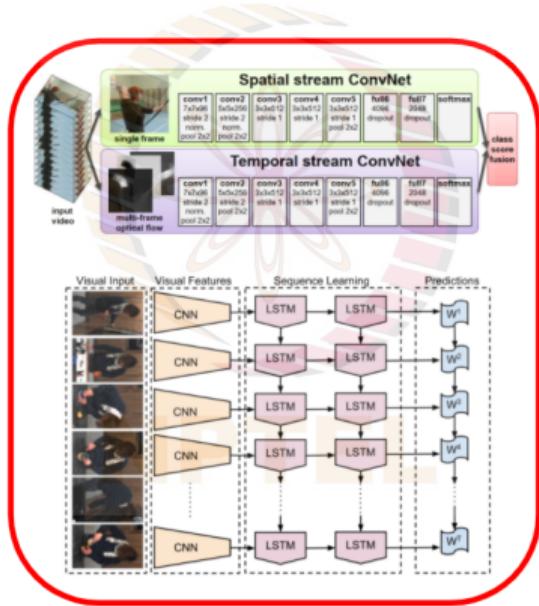


Action Recognition

Baseball Pitch, Basketball Shooting, Bench Press, Biking, Billiards Shot, Breaststroke, Clean and Jerk, Diving, Drunning, Fencing, Golf Swing, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jumping Jack, Jump Rope, Kayaking, Lunges, Military Parade, Mixing Batter, Nun chucks, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skate Boarding, Skiing, Skijet, Soccer Juggling, Swing, TaiChi, Tennis Swing, Throw Discus, Trampoline Jumping, Volleyball Spiking, Walking with a dog, Yo Yo
Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Basketball Dunk, Blow Drying Hair, Blowing Candles, Body Weight Squats, Boweling, Boxing-Punching Bag, Boxing-Speed Bag, Brushing Teeth, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, From Crawl, Hair cut, Hammering, Hammer Throw, Handstand Pushups, Handstand Walking, Head Massage, Ice Dancing, Knitting, Long Jump, Mopping Floor, Parallel Bars, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Rafting, Shaving Beard, Shot put, Sky Diving, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Table Tennis Shot, Typing, Uneven Bars, Wall Pushups, Writing On Board

Soomro et al, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, 2012

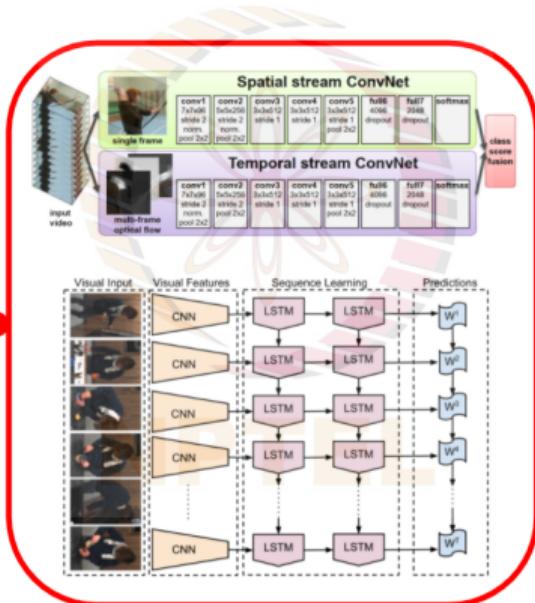
What all can be done?



Sigurdsson et al, Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, ECCV 2016

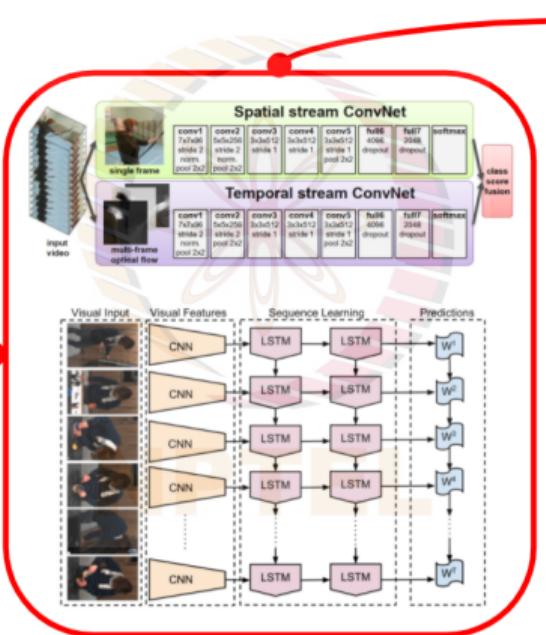
What all can be done?

Train - Hollywood in Homes



What all can be done?

Train - Hollywood in Homes

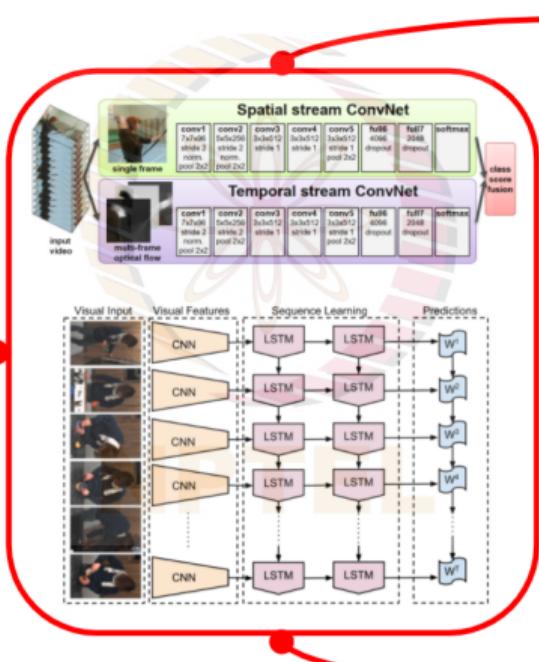


Action Recognition

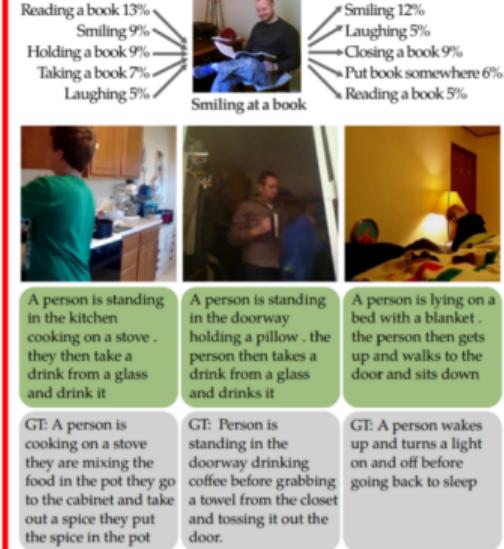


What all can be done?

Train - Hollywood in Homes



Action Recognition



Sentence Prediction

Sigurdsson et al, Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, ECCV 2016

Other Tasks in Video Understanding

- Action Forecasting
- Object Tracking
- Dynamic scene understanding
- Temporal Action Segmentation
- ...



Homework

Readings

- Tutorial on Large-scale Holistic Video Understanding
- <https://paperswithcode.com/area/computer-vision/video>

Question

- What do you think will happen if you train a model on normal videos and do inference on a reversed video?