Video Understanding using Deep Learning Algorithms

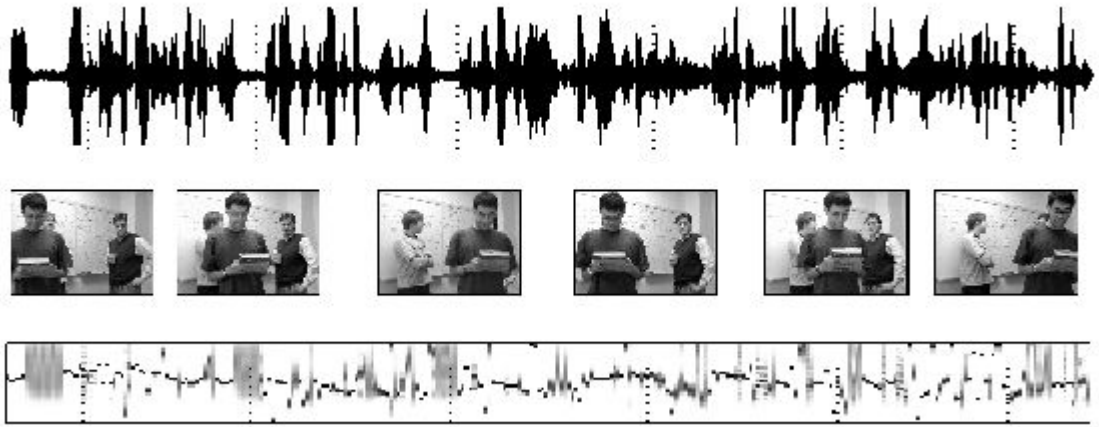
Masoud Kaviani
Senior Data Scientist at SabaIdea (Aparat, Filimo, Cinematicket)

What is a Video

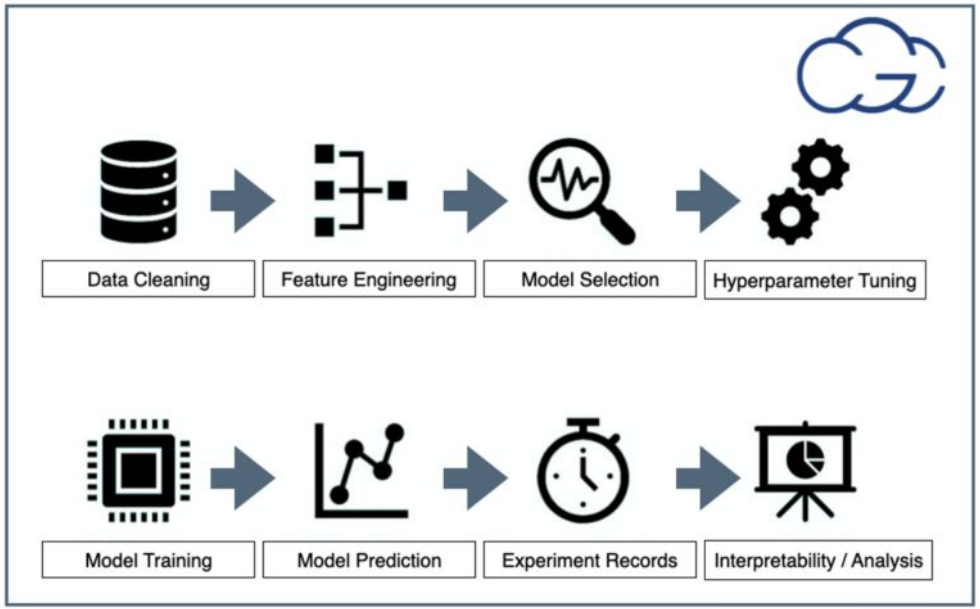
Audio

+

Sequence of Images

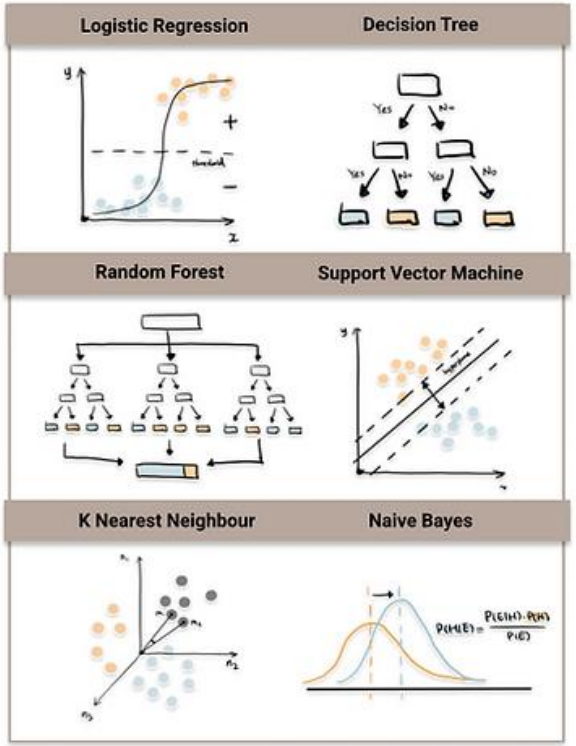


Machine Learning

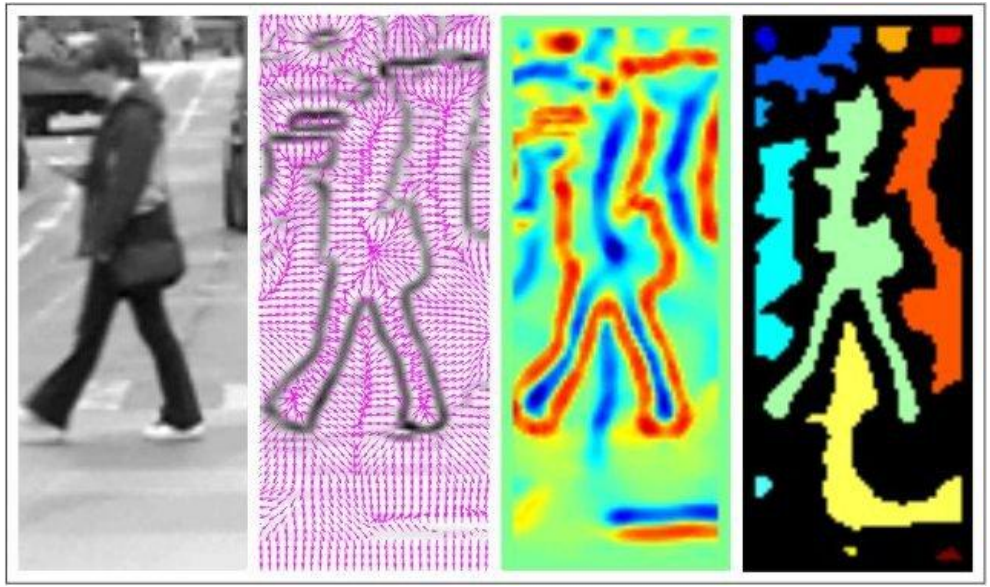


Old School Algorithms

- Simple Pattern in Data
- Small Amount of Data

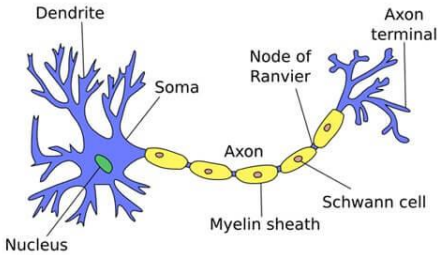


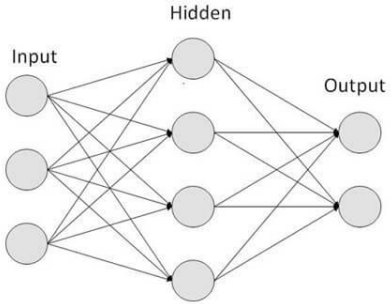
Pattern Recognition in Videos



Neural Networks

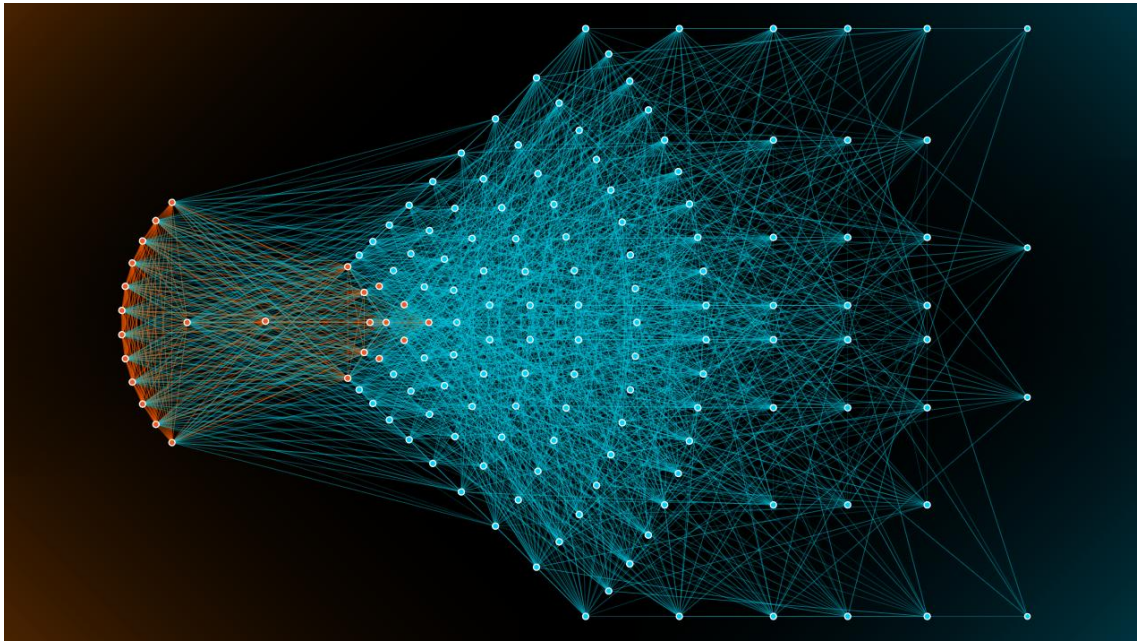
Simulate Learning Like Neurons inside Human Brains



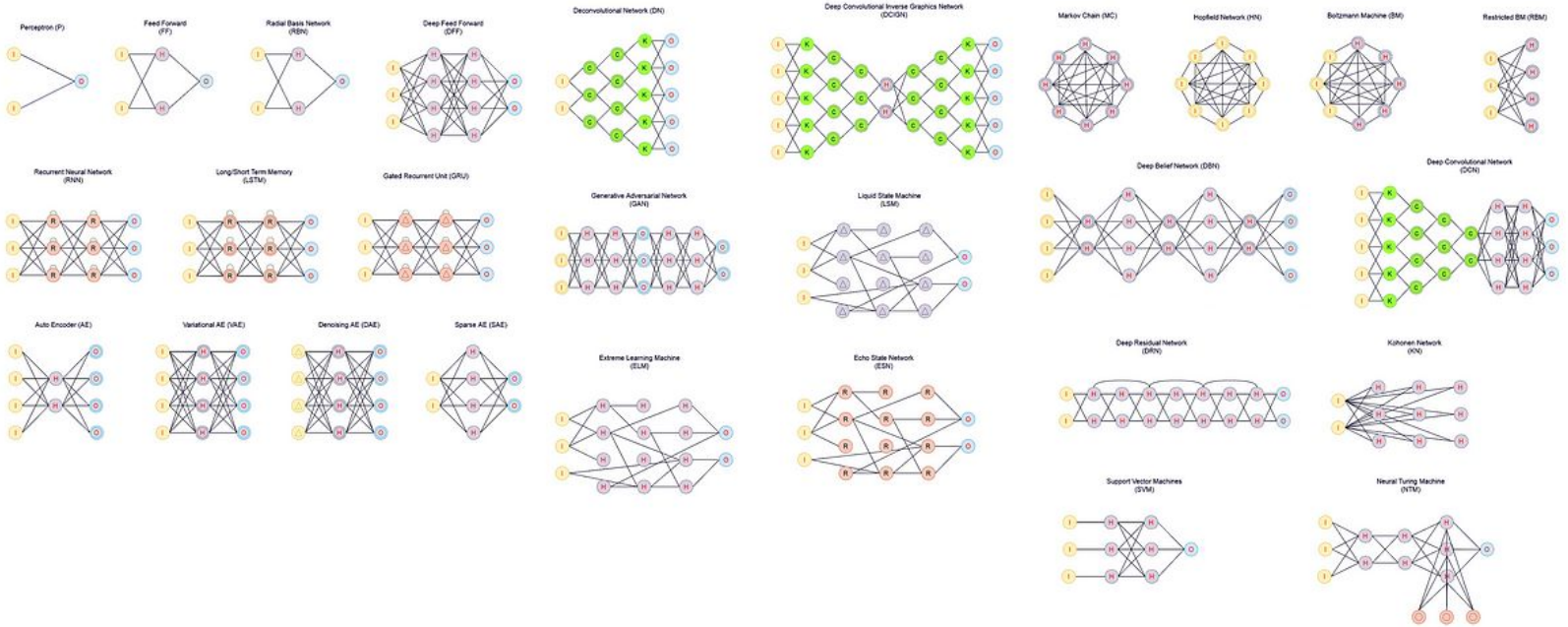


Deep Neural Networks

- Learn Complex Patterns
- Large Amount of Data

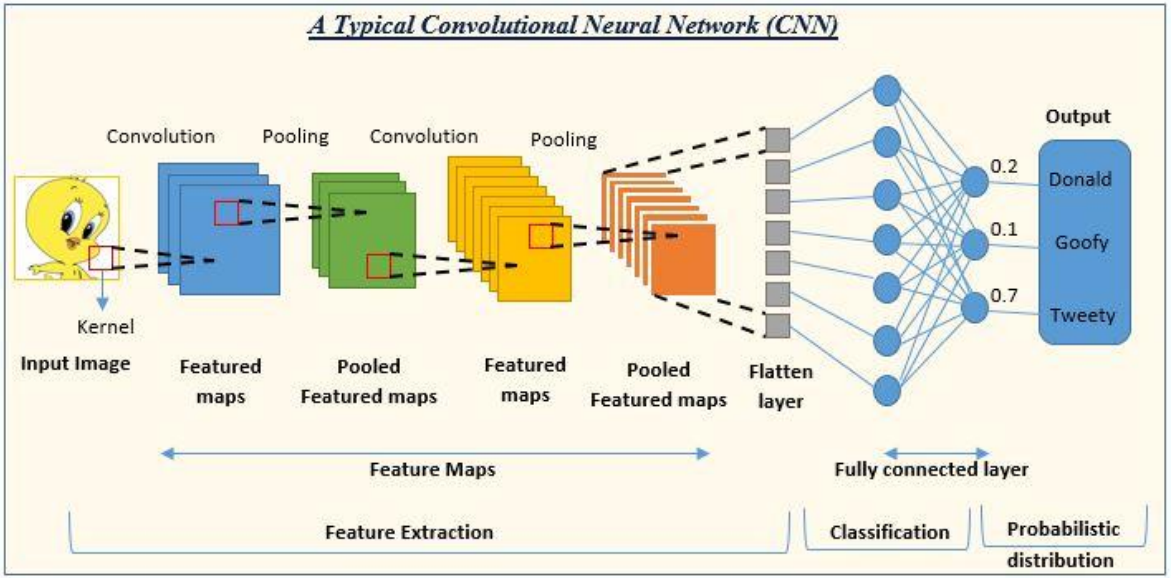


Different Types of Deep Neural Networks



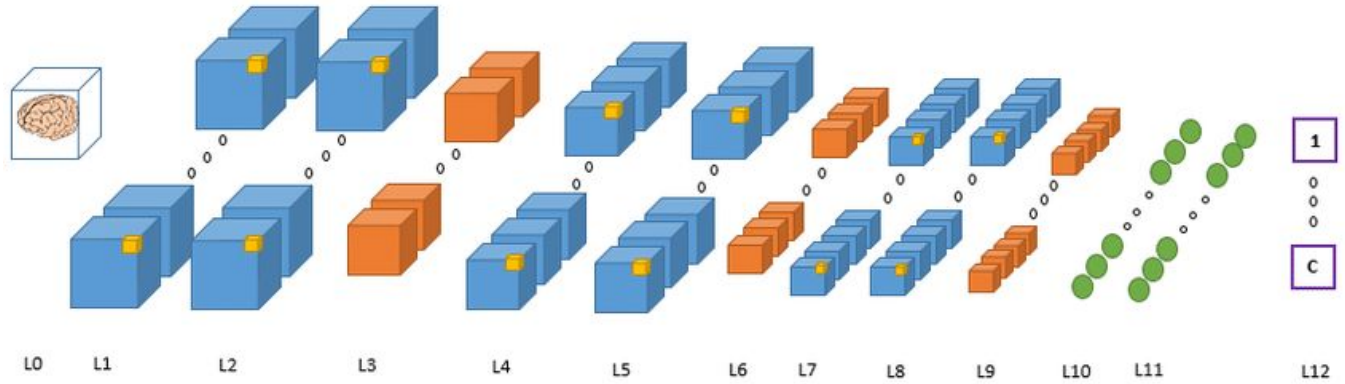
Convolutional Neural Networks

- Very Good for Images
- Using Convolutions and Pooling



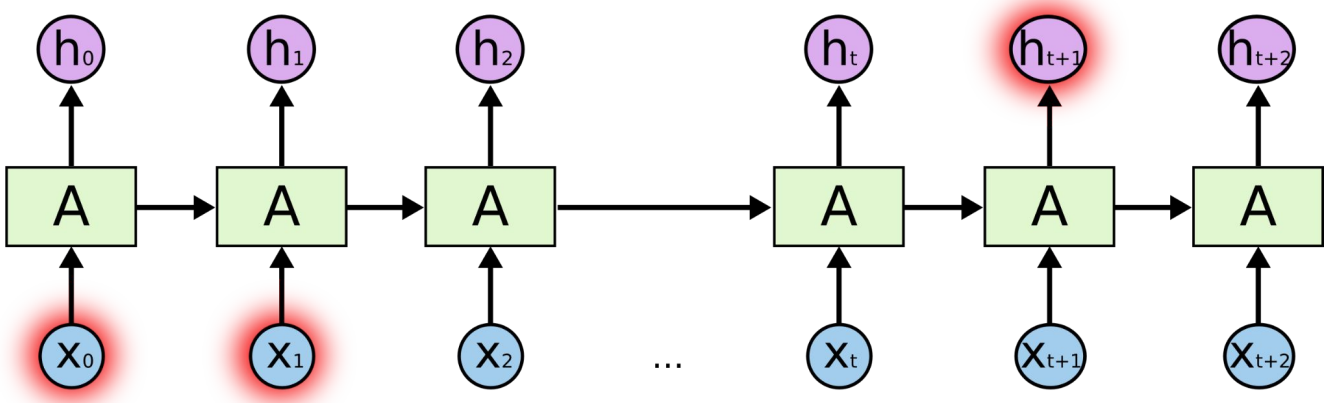
3D Convolutional Neural Networks

- Better Performance but Higher Computation



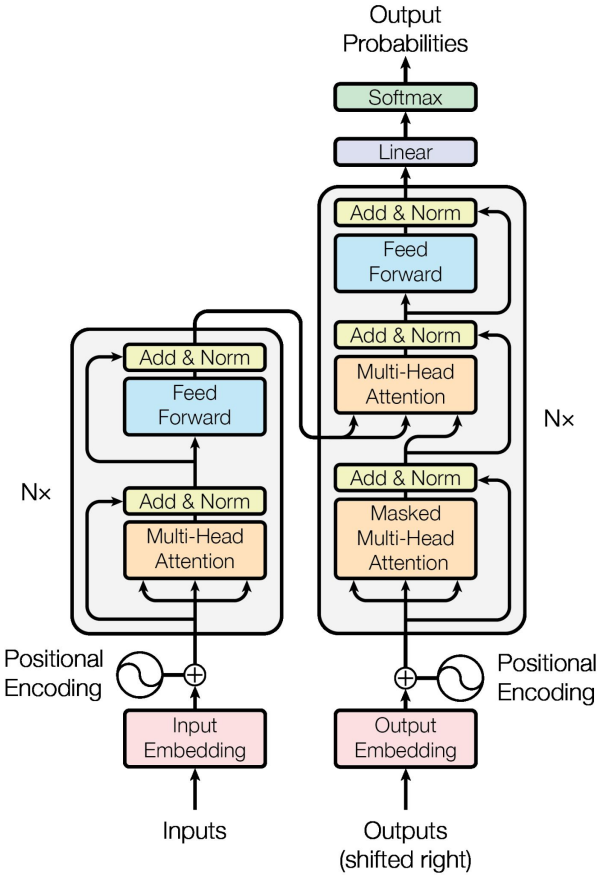
RNN/LSTM/GRU

Recurrent Neural Networks → Sequence of Data



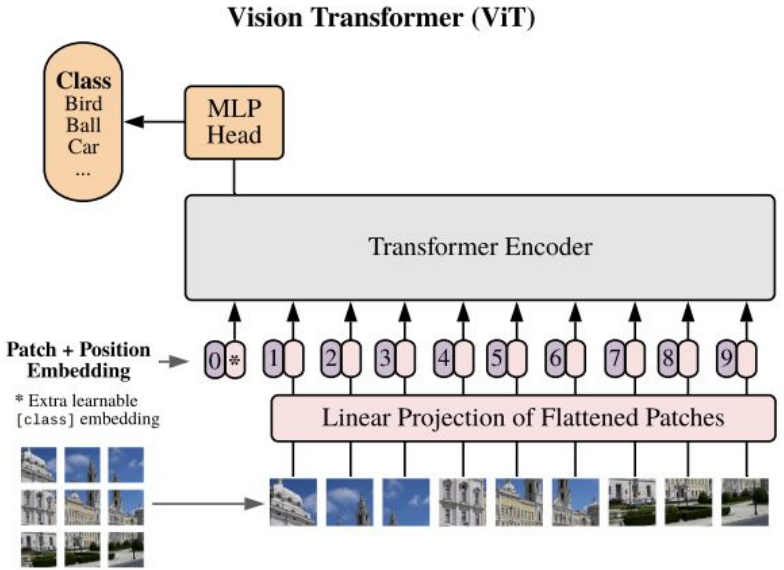
Attention is All you Need

- Handling Seq2Seq
- Long-range Dependency
- Using Parallel Processing



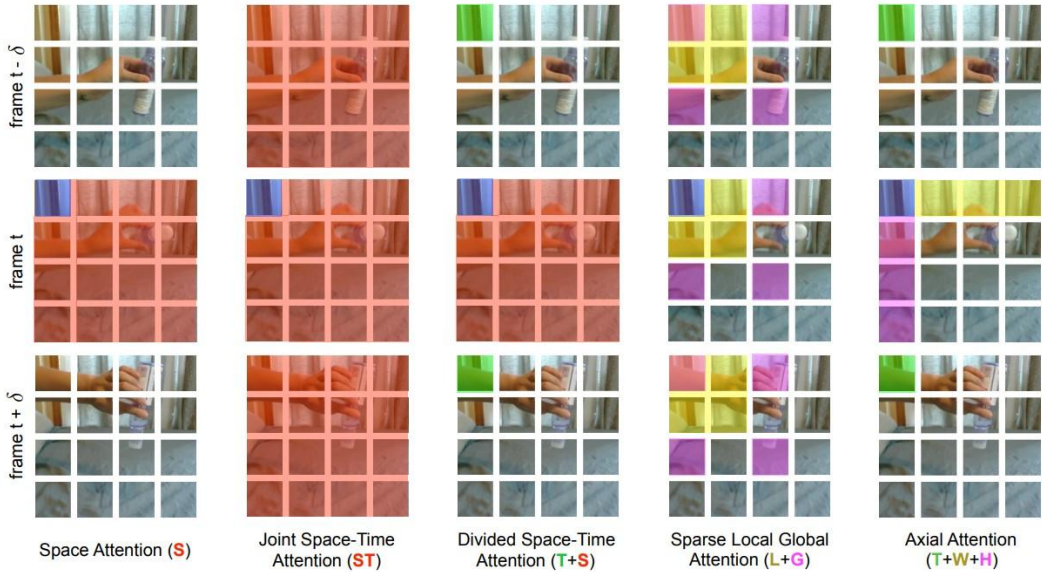
An Image Worth 16x16 Words

- Extract Information from Images by Patches
- Parallel Processing
- Using Transformers



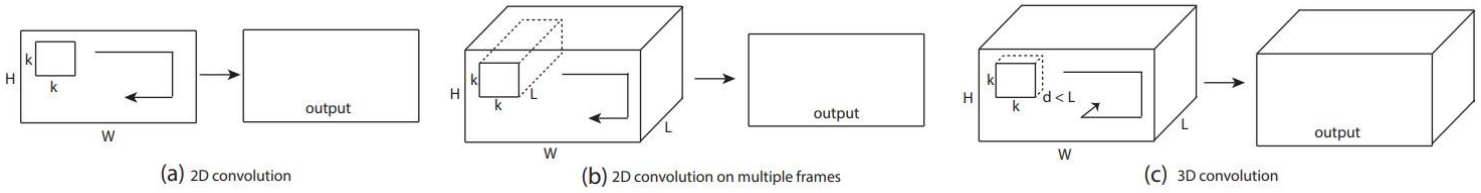
Is Space-Time Attention All You Need for Video Understanding?

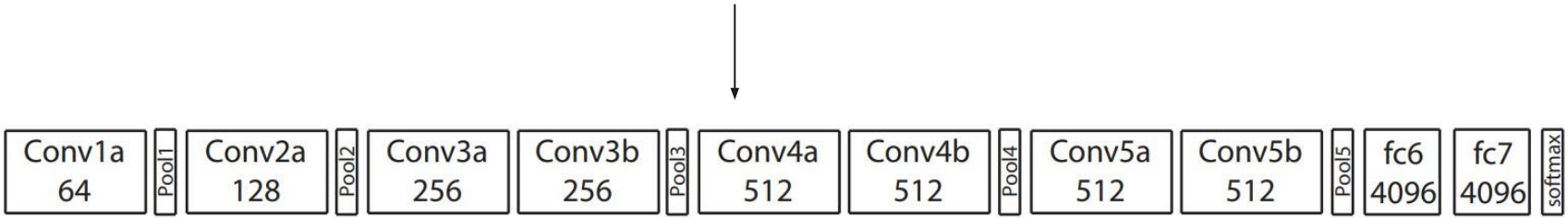
- Using Attention Mechanism to Understand Video for Video Classification
- Better Performance than 3D Convolutional



State-of-the-Art Models

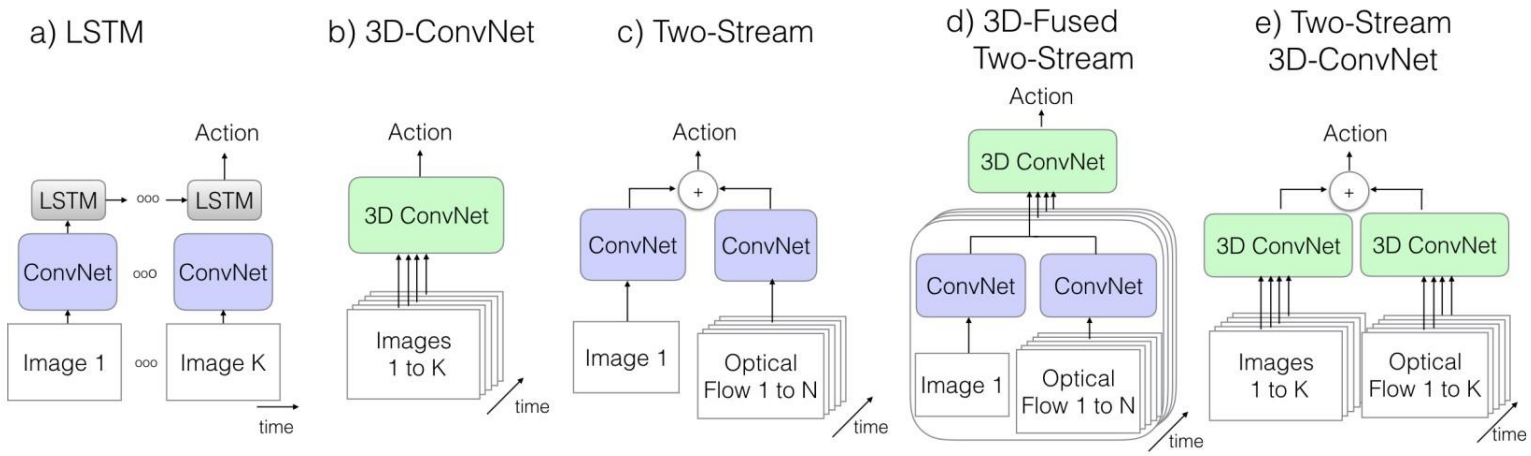
- Learning Spatiotemporal Features with 3D Convolutional Networks





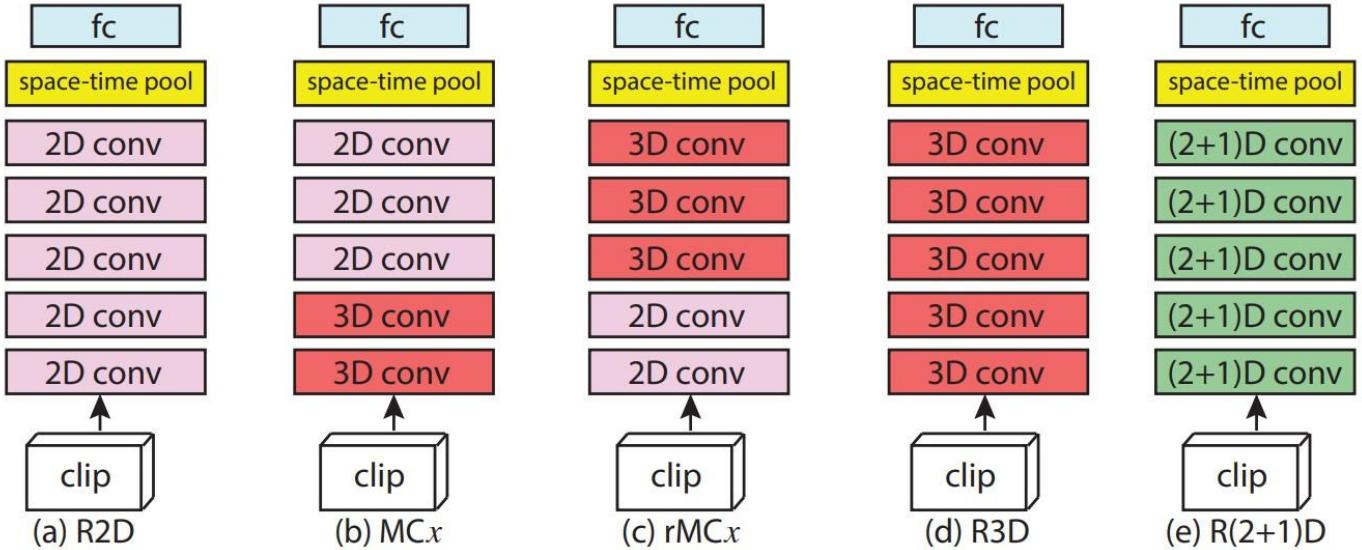
State-of-the-Art Models

- Two-Stream Inflated 3D ConvNets



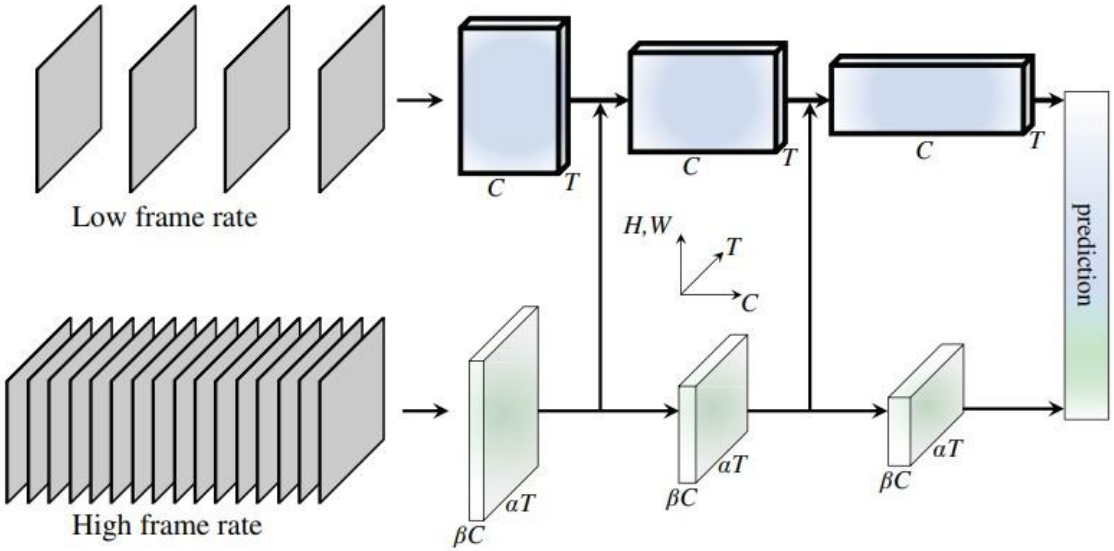
State-of-the-Art Models

- R2+1D



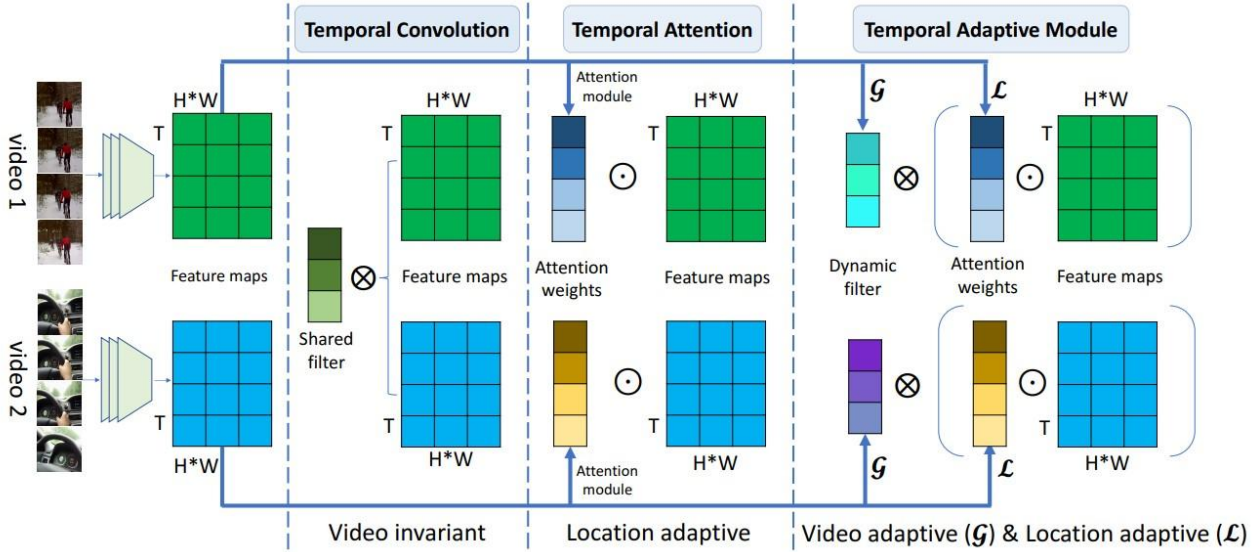
State-of-the-Art Models

- SlowFast Networks



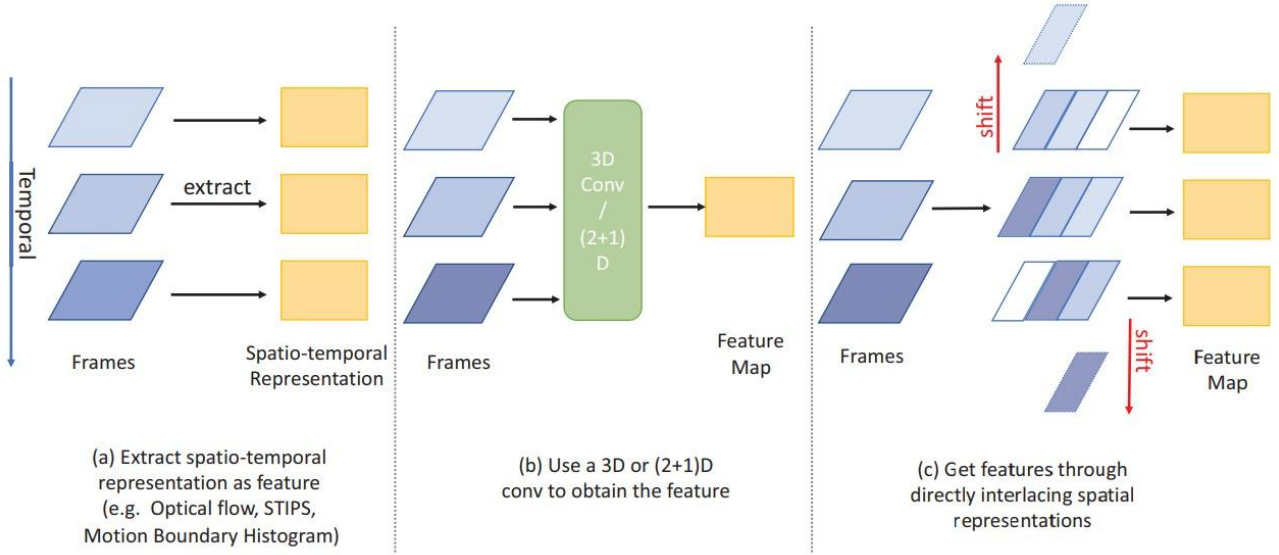
State-of-the-Art Models

- TAM: Temporal Adaptive Module for Video Recognition



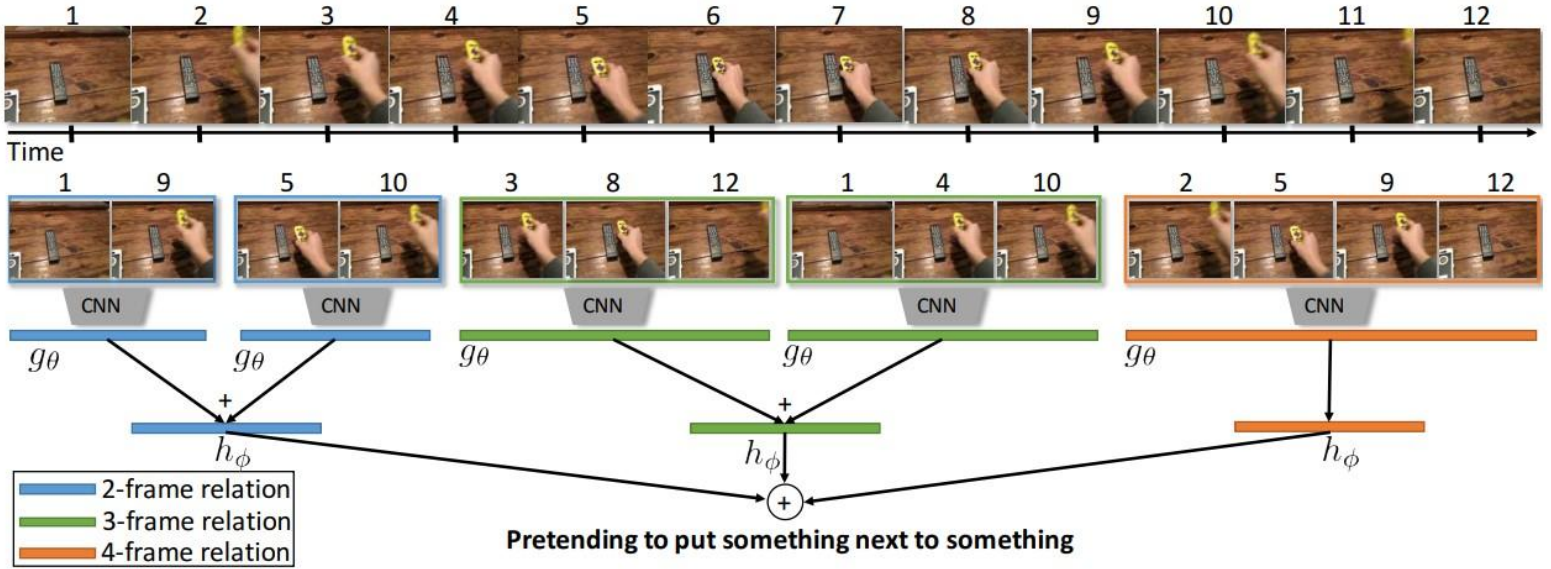
State-of-the-Art Models

- Temporal Interlacing Network



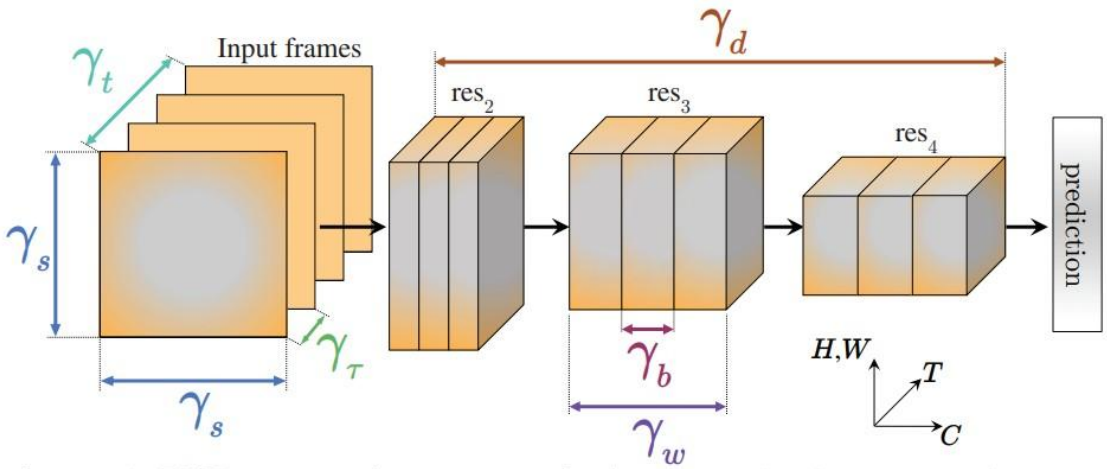
State-of-the-Art Models

- Temporal Relational Reasoning in Videos



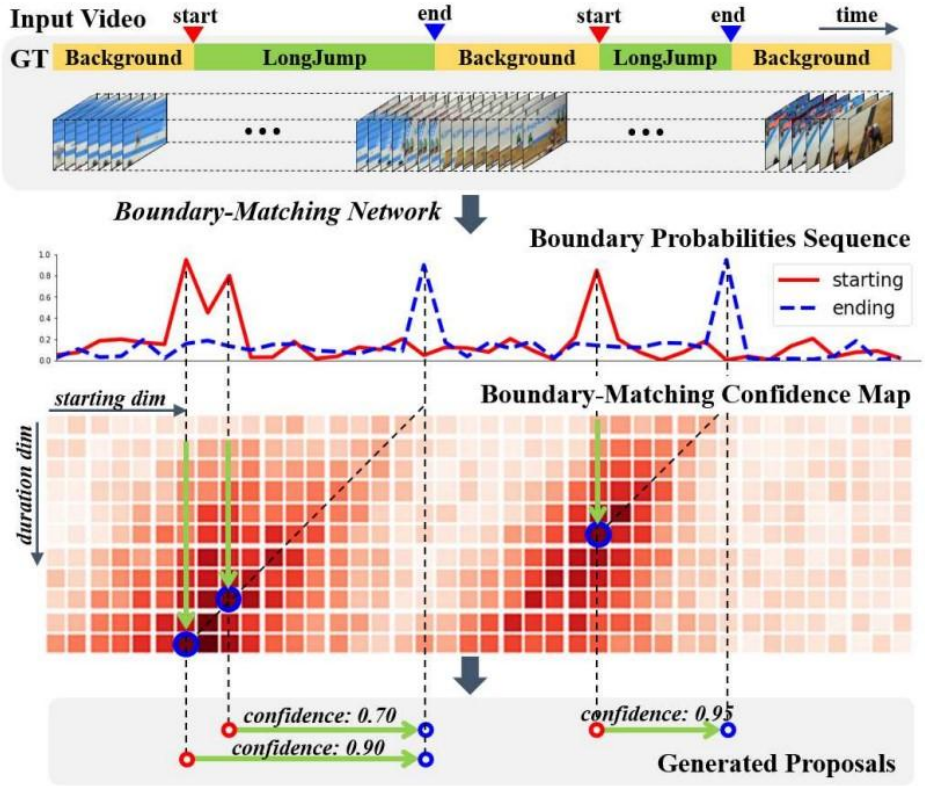
State-of-the-Art Models

- X3D: Expanding Architectures for Efficient Video Recognition



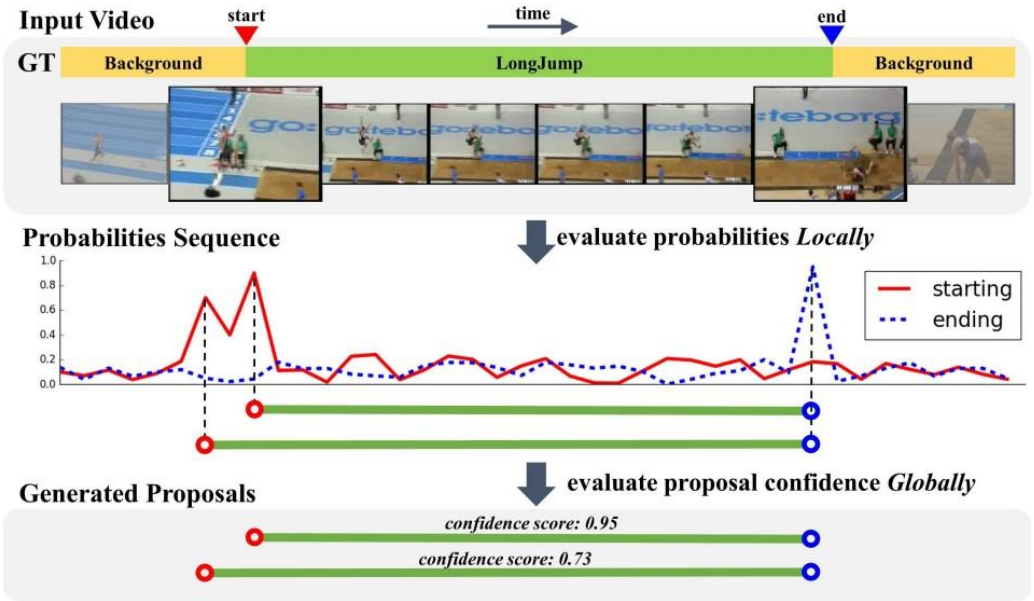
State-of-the-Art Models

- BMN: Boundary-Matching Network for Temporal Action Proposal Generation



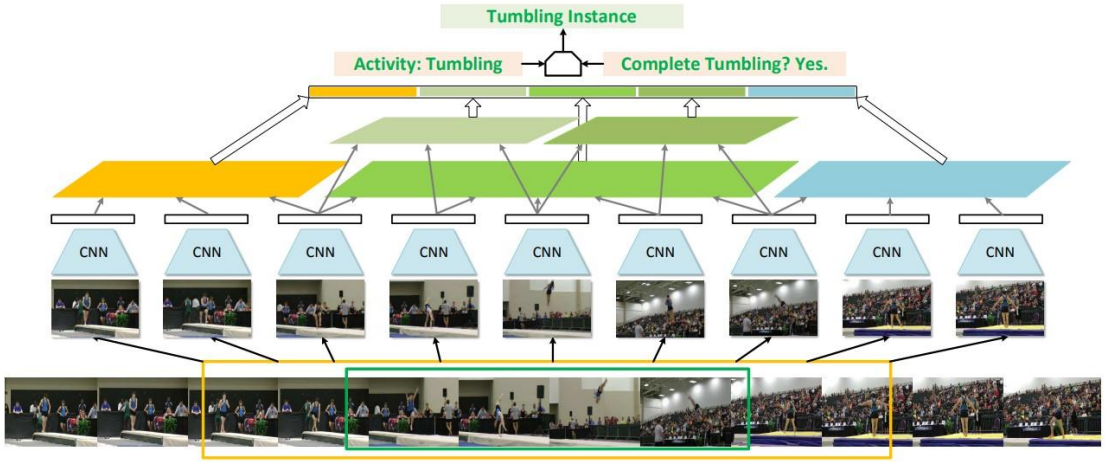
State-of-the-Art Models

- BSN: Boundary Sensitive Network for Temporal Action Proposal Generation



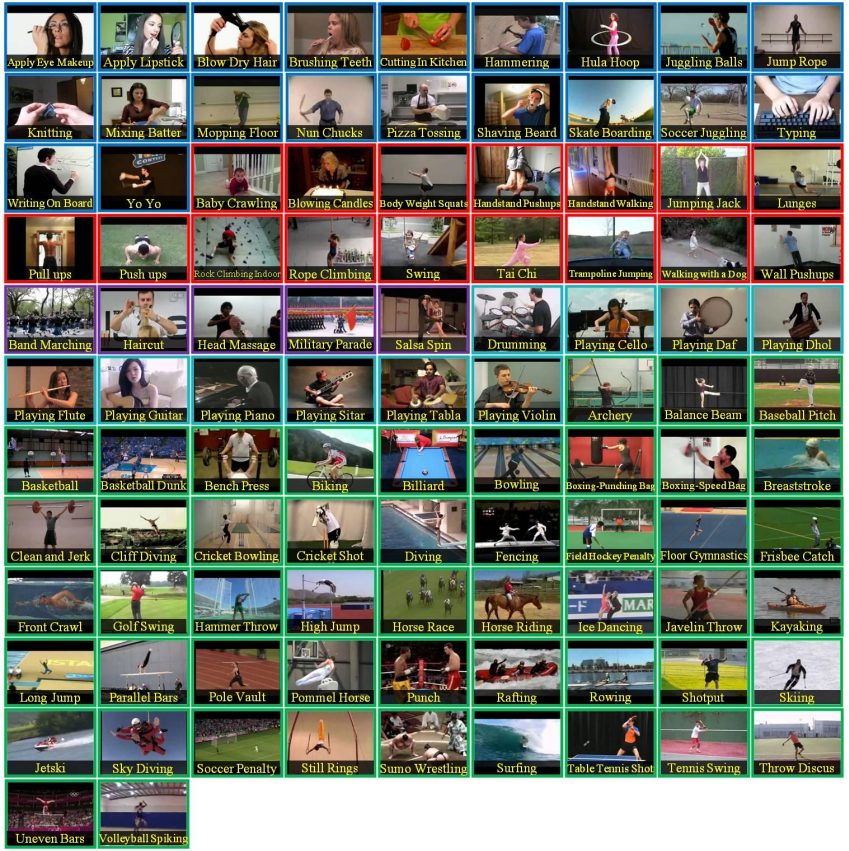
State-of-the-Art Models

- Temporal Action Detection With Structured Segment Networks



DataSets

- UCF101 Dataset (www.crcv.ucf.edu/data/UCF101.php)
- 101 Category Action from Youtube Videos



DataSets

- Kinetics 700 (paperswithcode.com/dataset/kinetics-700)
- 700 Human Actions

DataSets

- Holistic Video Understanding (holistic-video-understanding.github.io)
- 3142 Labels with Large Amount of Data

forest,musician,flutist,music,musical_instrument,brass_instrument,wind_instrument,flautist,recreation,musical_instrument_accessory,plant,playing_flute,tree

string_instrument,musician,man,,guitarist,plucked_string_instruments,music,tapping_guitar,bass,musical_instrument_accessory,performance,string_instrument_accessory,electric_guitar,sitting,monochrome_photography,musical_instrument,guitar_accessory,resonator

sport_venue,shoe,outdoor_shoe,joint,foot,ball,grass,knee,human_leg,fun,football_player,ball_game,green,footwear,football,player,sports_equipment,juggling_soccer_ball,soccer,plant,soccer_ball,sports,play

opening_bottle_not_wine_,joint,muscle,service,finger,distilled_beverage,fun,taste,standing,arm,t_shirt,glass,alcohol,drink,hand,bottle,photograph,cooking

smile,nose,textile,cheek,thigh,mouth,girl,diaper,finger,baby_products,human_leg,fun,playing_xylophone,infant,toy,facial_expression,skin,child,hand,sitting,human_hair_color,daytime,play,toddler

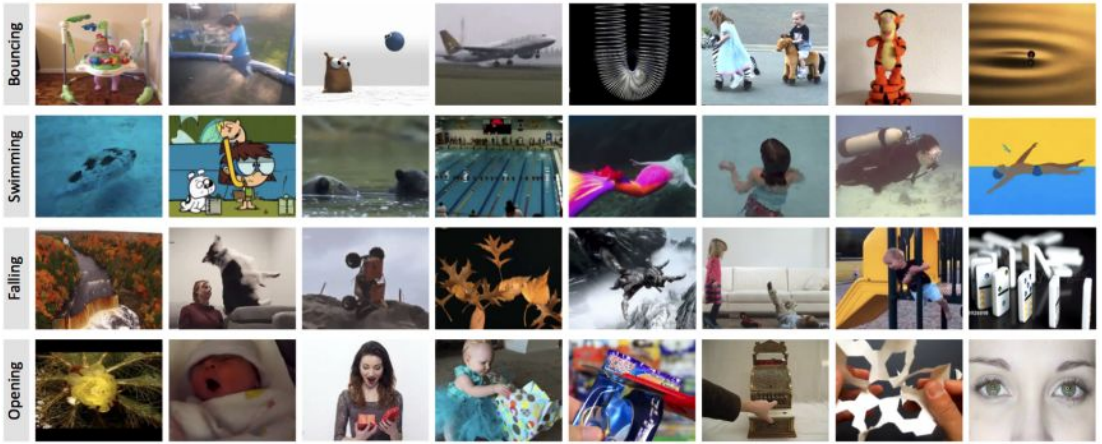
coast,watercourse,plant,wetland,terrain,floodplain,marsh,wading_through_mud,boulder,tree,water,natural_resources,river,rock,waterway,outcrop,shore,creek

DataSets

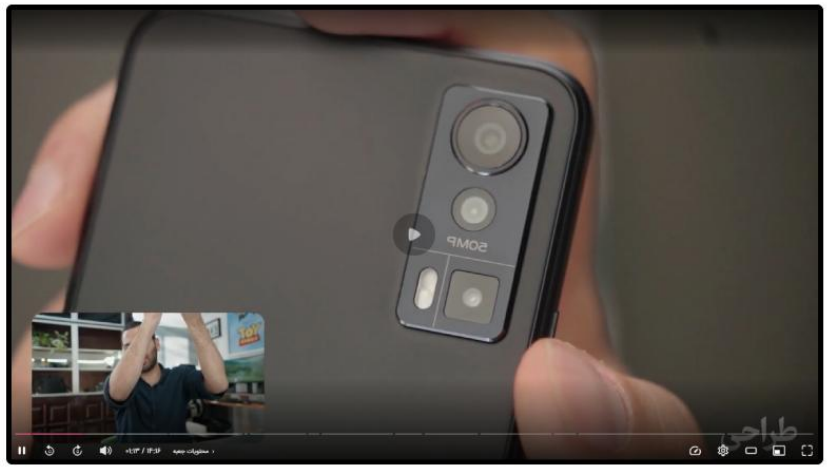
- Activity-Net (activity-net.org)
- 200 Activity Classes

DataSets

- Moments in Time (moments.csail.mit.edu)



A Simple Use Case



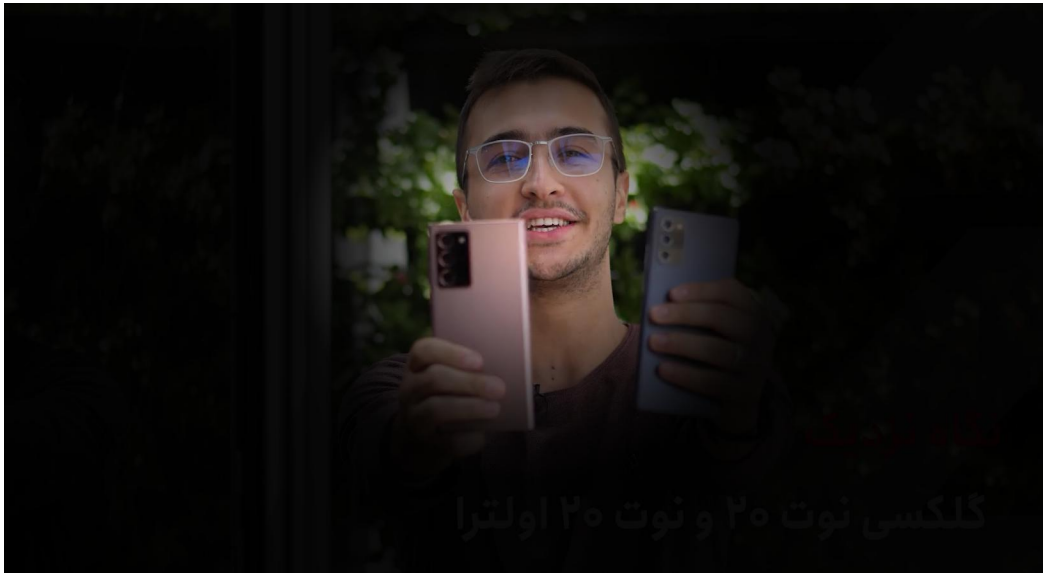
Split Video into Scenes

A Simple Use Case

A new video start processing

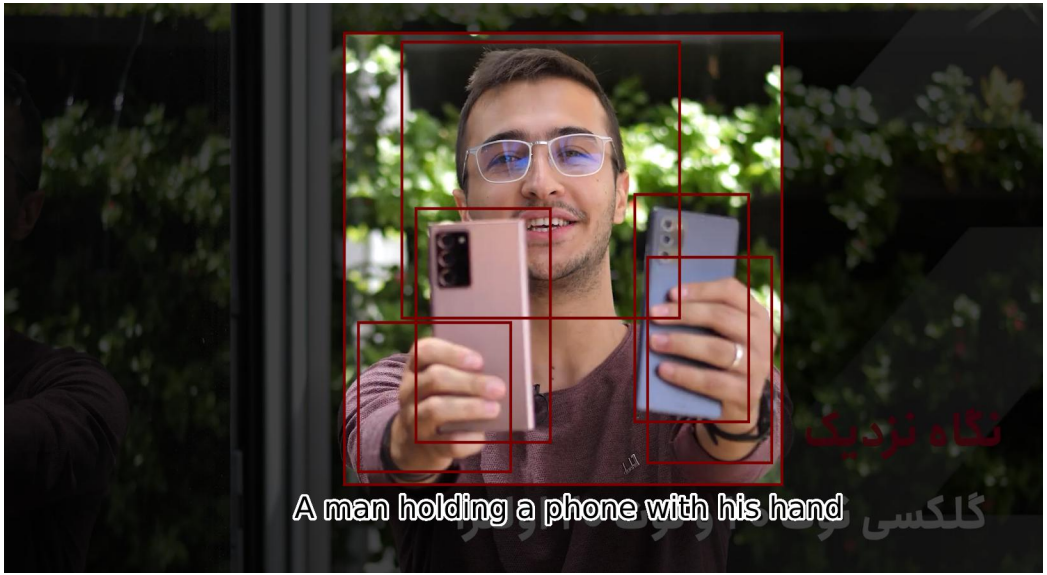
A Simple Use Case

Detect Interesting Locations



A Simple Use Case

Detect Objects at
Interesting Locations and Use them
to Understand Actions



Another Simple Use Case

A new video start processing

Another Simple Use Case

Detect Interesting Locations

Another Simple Use Case

Detect Objects at
Interesting Locations and Use them
to Understand Actions

Thank you for your Attention

You can find me on:

website: **masoudkaviani.ir**

twitter: **@masoud_kaviani**