REPUBLIC POLYTECHNIC
DISCOVER. TRANSFORM. ACHIEVE

Lesson 4: Human Activity Recognition

https://youtu.be/o7MAr8U4Un0

HEALTH DECLARATION

Please complete this mandatory declaration

Together, let us stay vigilant, stay healthy and look after one another

Programme (Day 2)	
Section 1:	What is Human Activity Recognition?
Section 2:	What is Sequence Data?
Section 3:	Publicly Available Human Activity Datasets
Section 4:	Neural Networks (Brief)
Tea Break	
Section 5:	Human Activity Recognition
Section 6:	Architectures for Action Recognition I
Section 7:	Lab1: Action Recognition with Two-Stream Inflated 3D ConvNet (I3D)
Lunch Break	
Section 8:	Architectures for Action Recognition II
Section 9:	Lab2: Action Recognition with MMAction2
Tea Break	
Section 10:	Lab3: Creating your own action recognition model
~ End ~	

What are you able to deduce from these series of events?

1. Two man running
2. One man holding up a machete (machete a.k.a. parang)
3. One man holding up a sword
4. One man running

Robbery gone wrong

https://youtu.be/aaMWyWskqSw

So Why Human Activity Recognition?

*500 hours of **video** are uploaded to **YouTube** every **minute**.

Streaming of videos 24/7

Impossible for manual annotation!

REPUBLIC POLYTECHNIC
DISCOVER. TRANSFORM. ACHIEVE

What is Human Activity Recognition?

Human Activity Recognition

- An automated understanding of a video interpreting human movement.
- In the example video given, you would want the video to be able to label and identify a robbery taking place.
- To be able to achieve this, contextual awareness of human activity needs to be applied to the video first in order for the system to automate the understanding and interpretation of the video.

Human Activity Recognition

- Contextual awareness is achieved through various level of human activities which can present in many different forms.
- A combination of a few human activities can then can come together to creating an understanding for a system:

Human-object interaction

Human-human interaction

Multi-body Part movement

Pose/Gesture

Human Activity Recognition

- Human activity recognition is achieved through training an action classifier in a supervised learning setup.
- The action classifier is created through the training of sequence data in order to produce an action recognition model.
- Common methods employed in the training of the models can include different variants of **Long short-term memory** (**LSTM**), which is also known as a **Recurrent Neural Network** (**RNN**).

What is Sequence Data?

Sequence Data

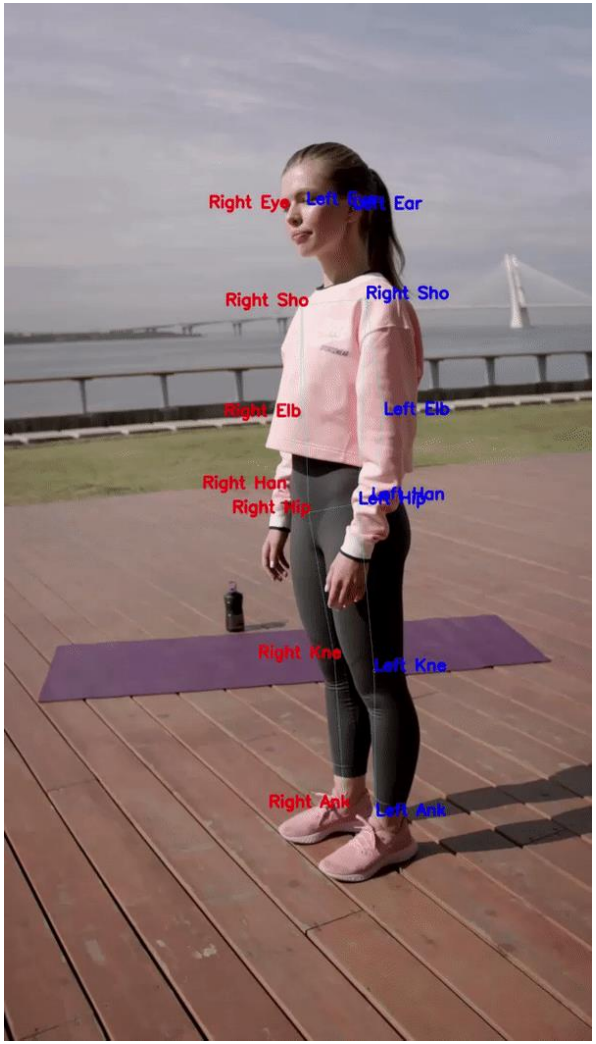
- Sequence Data, are data that occur in a series/sequence.
- Such data exist as a whole to create a meaning, while as an individua looses it.
- Example of sequential data can come as a form of text, where all the words come together to forms the meaning behind the intention, while as an individual looses it.

Sequence Data

- Another popular example will be stock prices where the prices of stocks are collated over time. This type of data are commonly referred to as time-series data. The analysis of such time-series data is commonly used for the purpose of forecasting stock prices and will prove meaningless if the relationship between time and prices are taken away during analysis.

Sequence Data

- The data that is employed by action recognition adopts the use of the human coordinates that are being identified in each frame of the video, time-series these data over the duration of the video/action.



Publicly Available Human Activity Datasets

UCF101

- UCF101: collection of videos from YouTube with 101 action categories.

https://youtu.be/FOajbVDTRIU

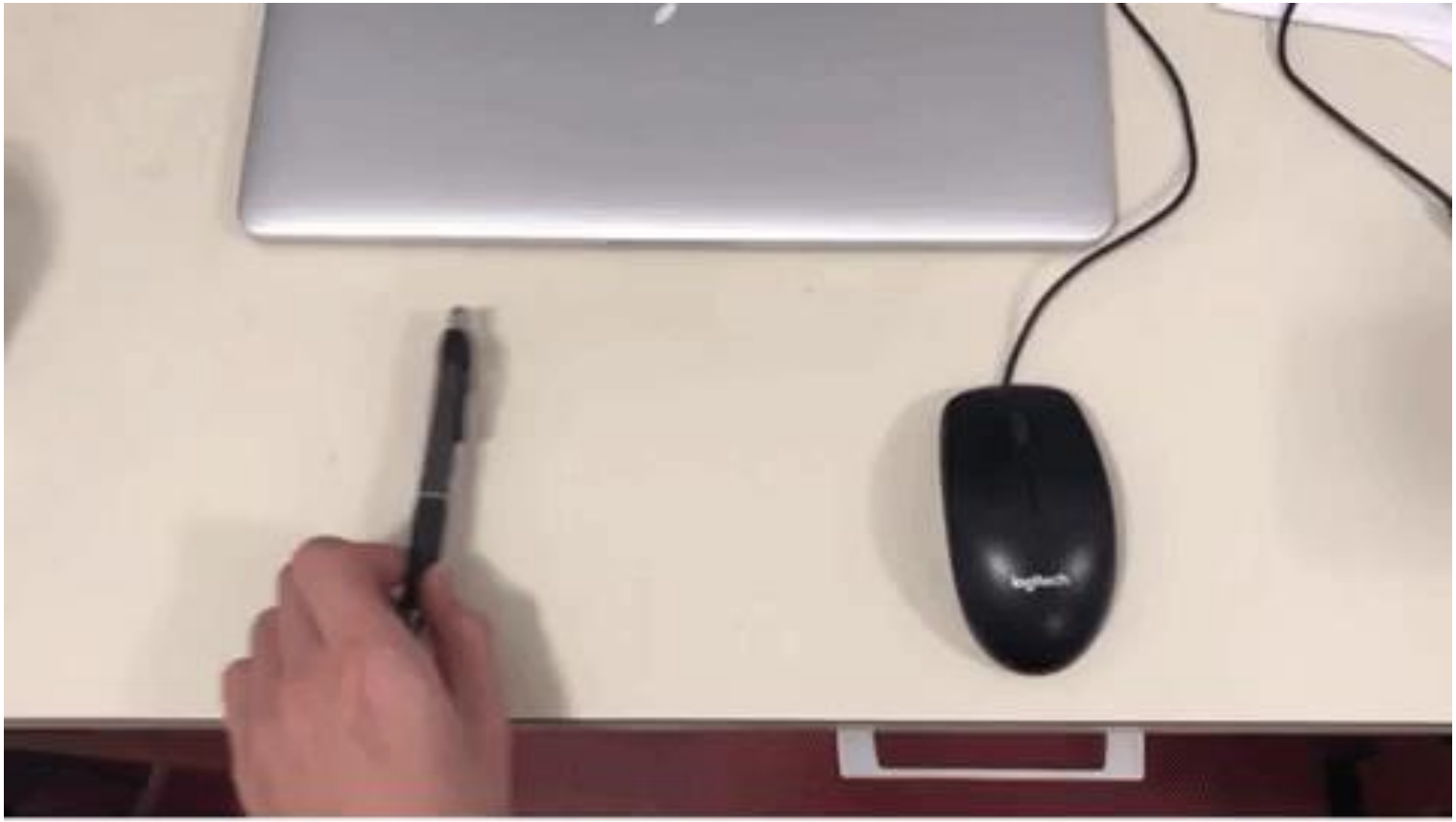
Sports-1M

- This dataset contains 1,133,158 video URLs which have been annotated automatically with 487 labels. The annotation was done via the YouTube Topics API.

https://youtu.be/qrzQ_AB1DZk

SOMETHING-SOMETHING

- A large collection of labeled video clips that show humans performing pre-defined basic actions with everyday objects.

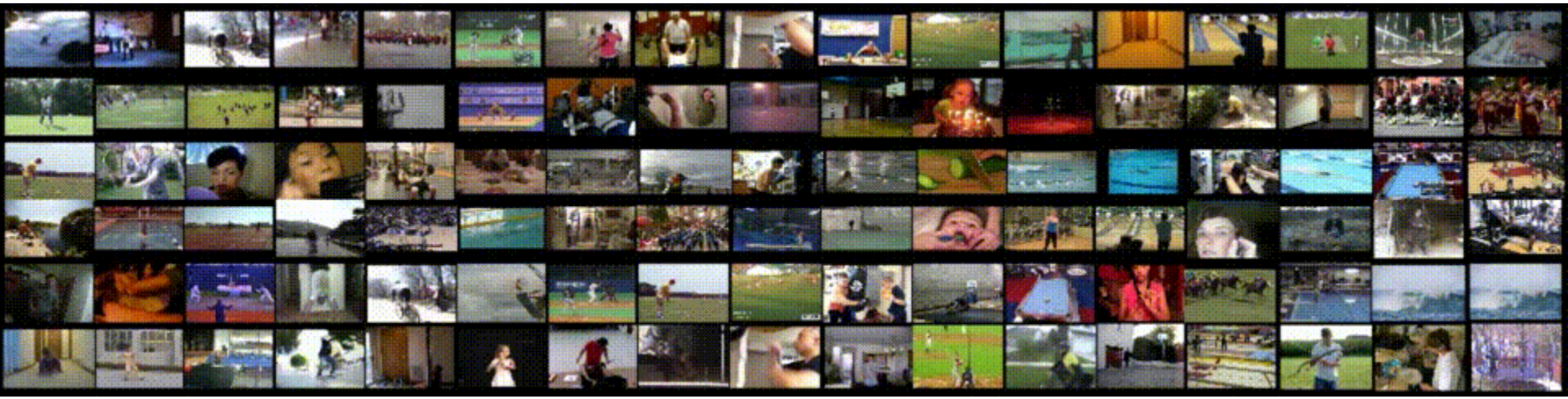


Prediction: Moving something closer to something

https://www.youtube.com/watch?v=C_ys1tDULpM

THUMOS 14

- A collection of over 254 hours of video data and 25 million frames split into various categories containing 101 actions. The dataset is offered with the purpose of action recognition for untrimmed videos.



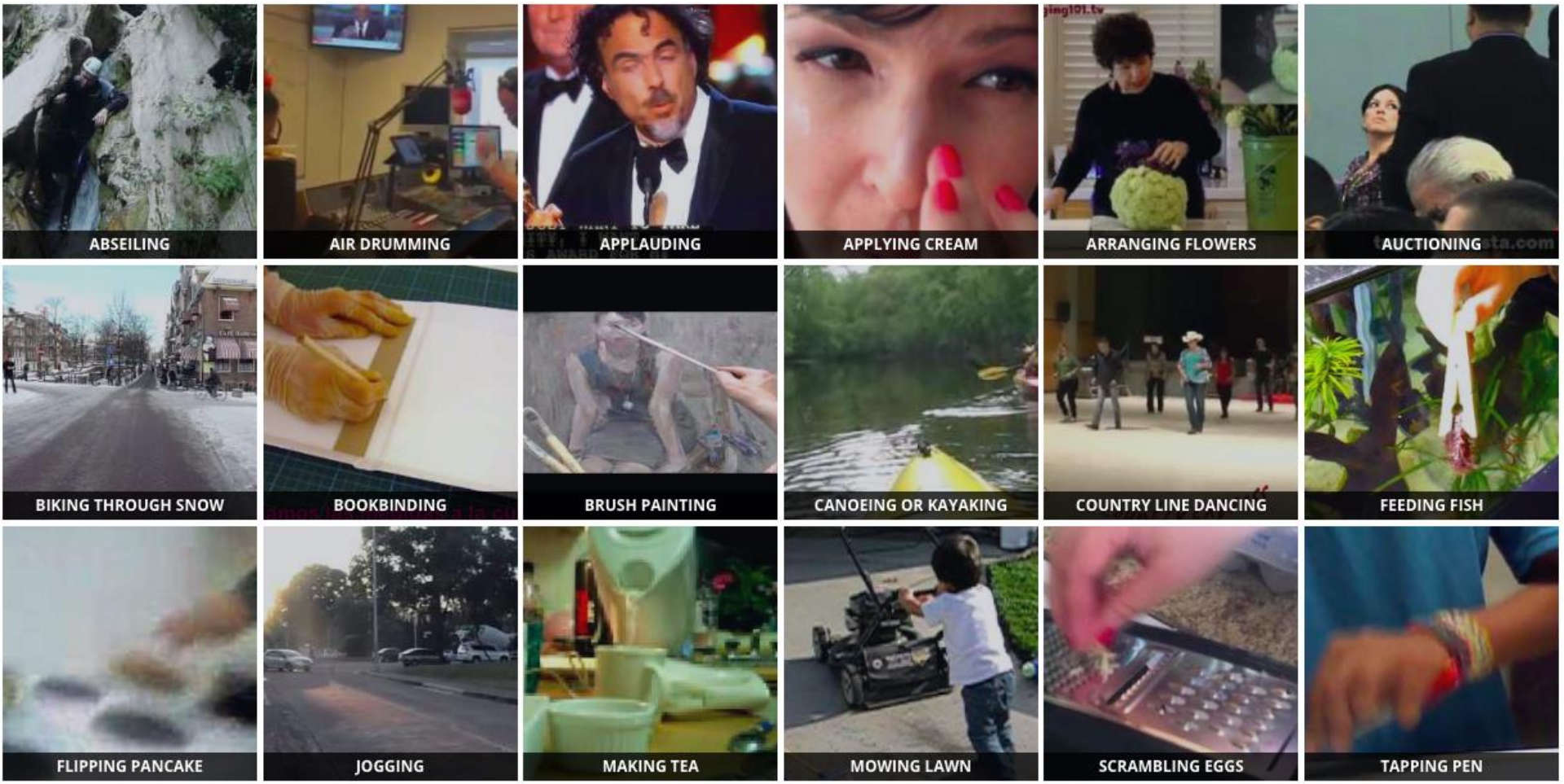
HMDB51

- HMDB: collection of videos from various sources, mostly movies with 51 action categories.

https://vimeo.com/141661915

Kinetics

- Kinetics: collection of video clips that cover up to 700 human action classes.



ActivityNet

- ActivityNet: collection of human activities that cover up to 200 human actions.

https://youtu.be/dYnEuvsRgeo

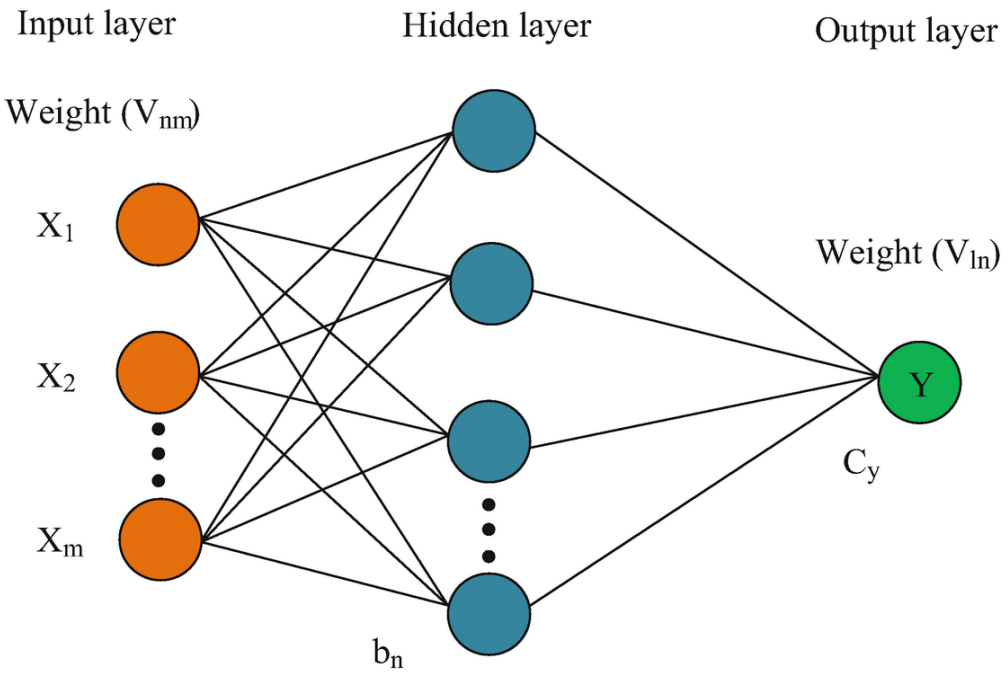
Neural Networks (Brief)

Neural Networks

- So which neural network should we be using?
- Before we take a look at the answer, let us take a minute to look through the different neural network below:
- Artificial Neural Network (ANN)
- Convolution Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Long short-term memory (LSTM)

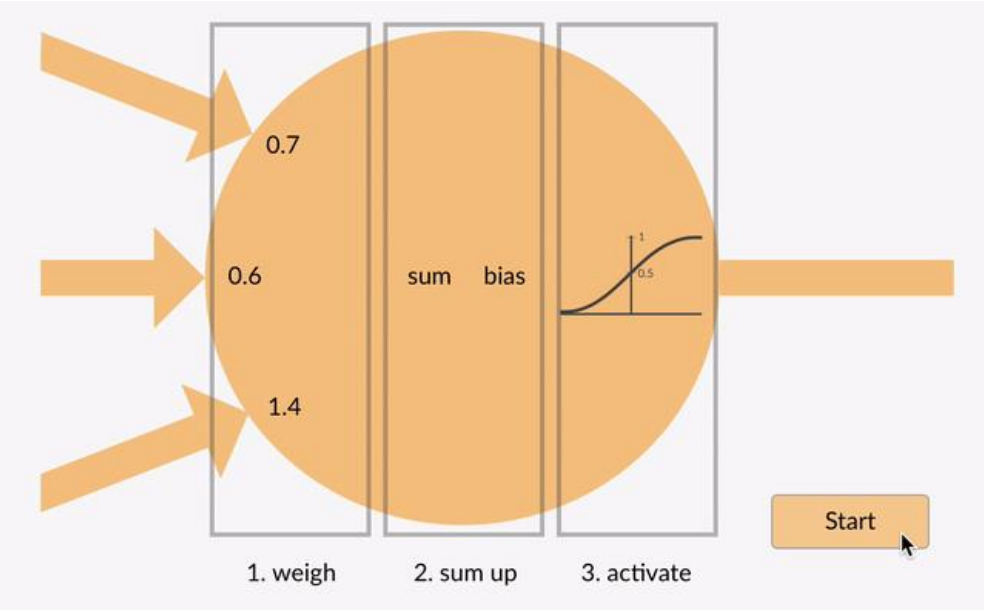
Artificial Neural Network (ANN)

- Artificial Neural Network (ANN) typically comprise of an input later, a hidden layer and an output layer.
- As how we process a single logistic regression, that can also be seen as a single neuron/perceptron in ANN.
- ANN is a feed forward network since data travels in a forward direction, from the input, through the hidden layers and to the output.



Artificial Neural Network (ANN)

- How a typical neuron in ANN works is by summing the weights together with the bias and passing it through an activation function.
- The activation function is much like an on/off switch where it will turn each feature on or off.
- This allows for the network to learn the relationship between the input and the output thereby solving complex problems.

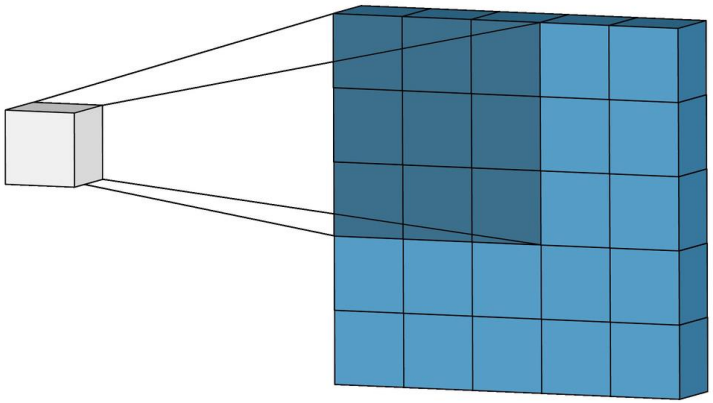


Artificial Neural Network (ANN)

- As good as ANN may be, it has its own set of challenges too.
- If used for image processing, it will require the conversion of a 2D image into a single dimension vector for processing.
- This would result in a loss of spatial features of an image.
- Spatial features refer to the arrangement of pixels and the relationship between them in an image.
- In addition to the above, ANN is also unable to capture sequential information in the input data which is required for dealing with sequence data.

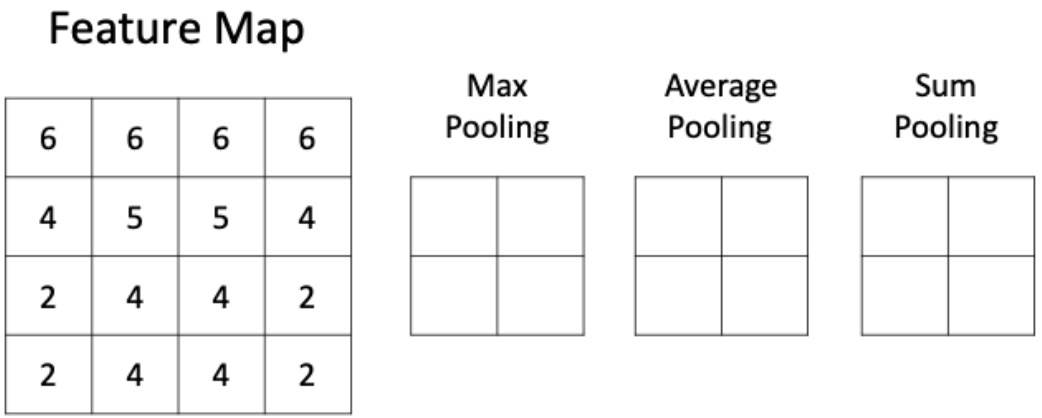
Convolution Neural Network (CNN)

- CNN makes use of kernels to extract the key features from the input using the convolution operation. By doing this, it is able to retain the spatial features from an image.
- The spatial feature will help us in identifying the object accurately, the location of an object, as well as its relation with other objects in an image.



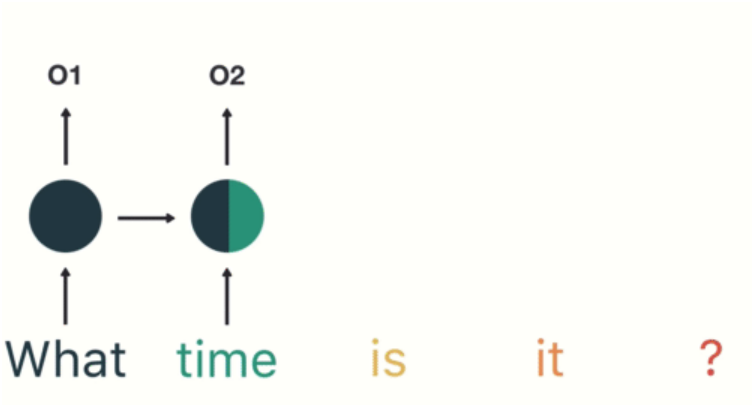
Pooling Layer

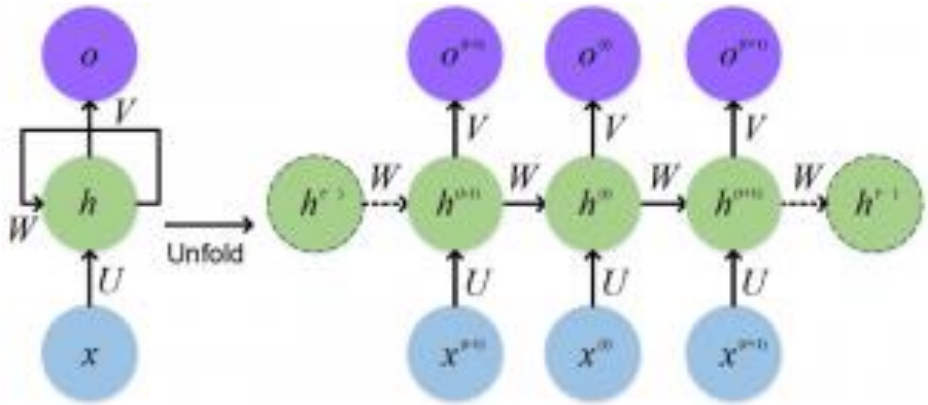
- When features are identified in the output, these features tend to have a bit more details than required (high resolution).
- Being too specific in terms of the identified feature may result in the feature map as being too rigid unable to be generalized for detection at other location of the image.
- In order to solve this problem, the feature maps are down sampled to allow the feature map to be reusable across the whole image.



Recurrent Neural Network (RNN)

- RNN is a class of artificial neural networks that allows information processed from one stage to flow to the next stage.
- This is achieved through looping the hidden layer as an input in the next layer.



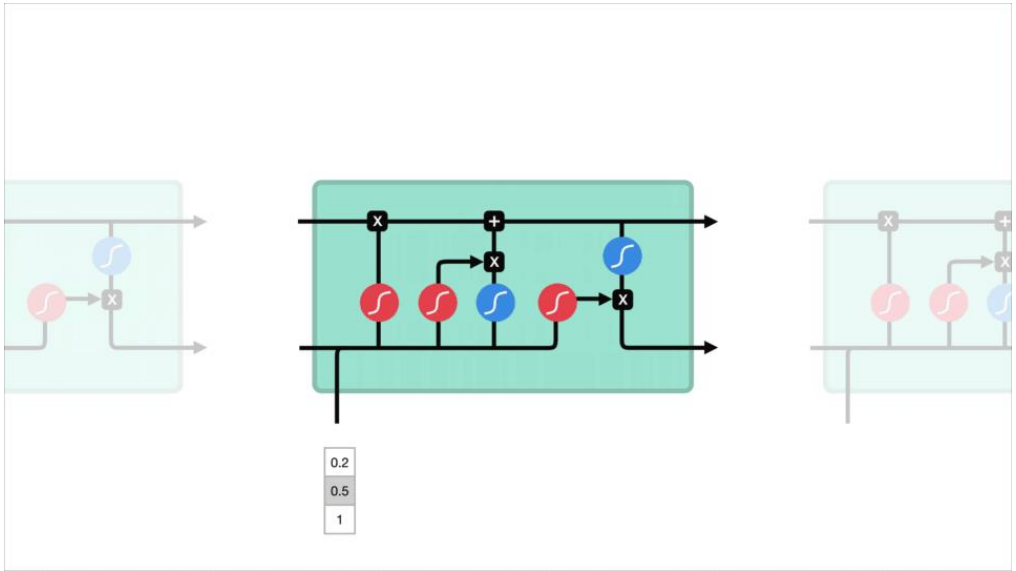


Recurrent Neural Network (RNN)

- The passing of information from previous stage as inputs allow the relationship between datapoints to be retained during processing.
- This is the reason why such networks are useful for time-series/sequential data analysis such as stocks prediction and natural language processing.
- In series data analysis, is it important that data processing does not loose the relationship within the data.

Long short-term memory (LSTM)

- LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning.
- The standard RNN lacks in terms of the ability to retain long terms information because as information is added for inference, the older data becomes less significant.
- As such, RNN was modified with gates to produce LSTM.
- What the gates literally do is to filter unnecessary information and to retain only important information.



How does all of these come together?

- With the use of CNN, we are able to achieve the spatial (location of the object with reference to the image/frame) knowledge of our object of interest.
- With the use of RNN, we are able to collate the knowledge behind the relationship between a sequence of frames.
- We will soon see how these two technologies come together to allow us to be able to extract the spatial-temporal information that allows us to understand human action/activity.

15 Mins Break

bit.ly/top10_2020

REPUBLIC POLYTECHNIC
DISCOVER. TRANSFORM. ACHIEVE

Human Activity Recognition

Human Activity Recognition

- Human activity recognition is achieved through a few different data extraction methods:
- *Wearable dependent data (accelerometer, gyroscope, magnetometer, and GPS of the smartphone)
- 3D method (mocap, Microsoft Kinect)
- 2D inference from video frames (much similar to pose recognition)

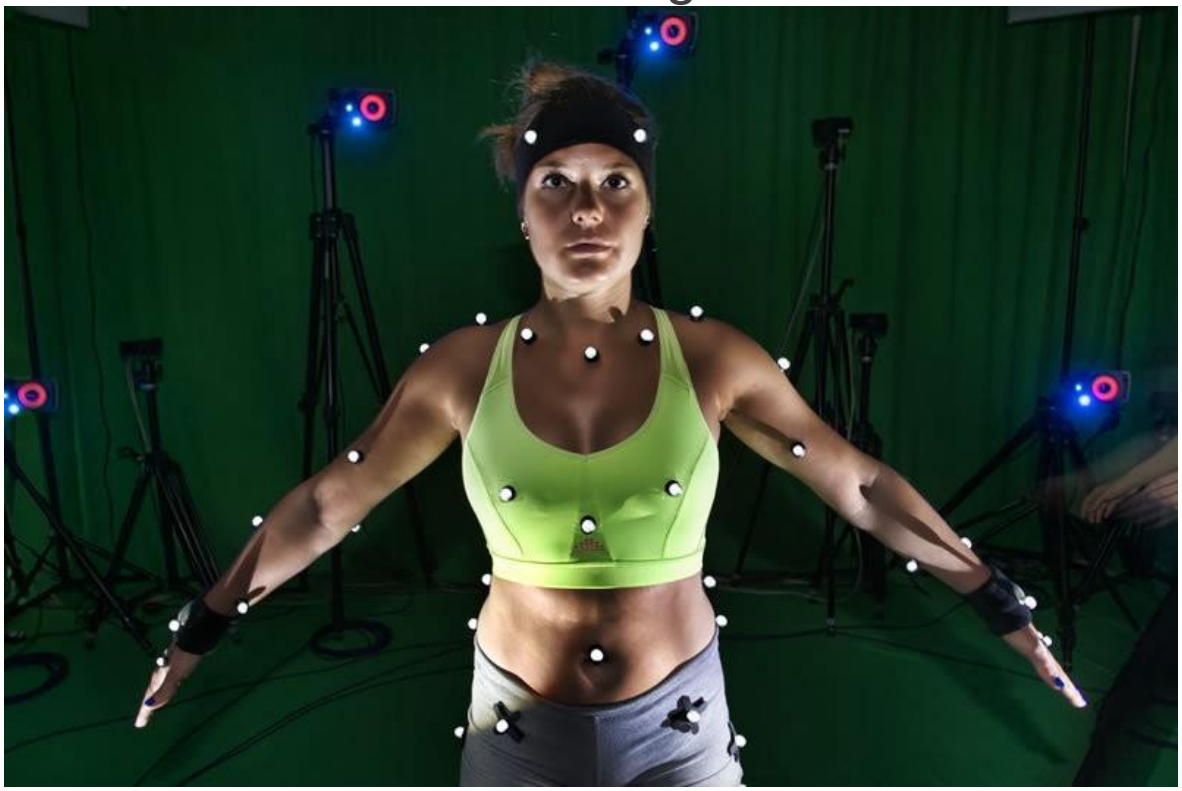
Human Activity Recognition

- Data collation through sensors can come from various sensors/hardware such as accelerometer, gyroscope, magnetometer, and GPS of the smartphone.

https://youtu.be/XOEN9W05_4A

Human Activity Recognition

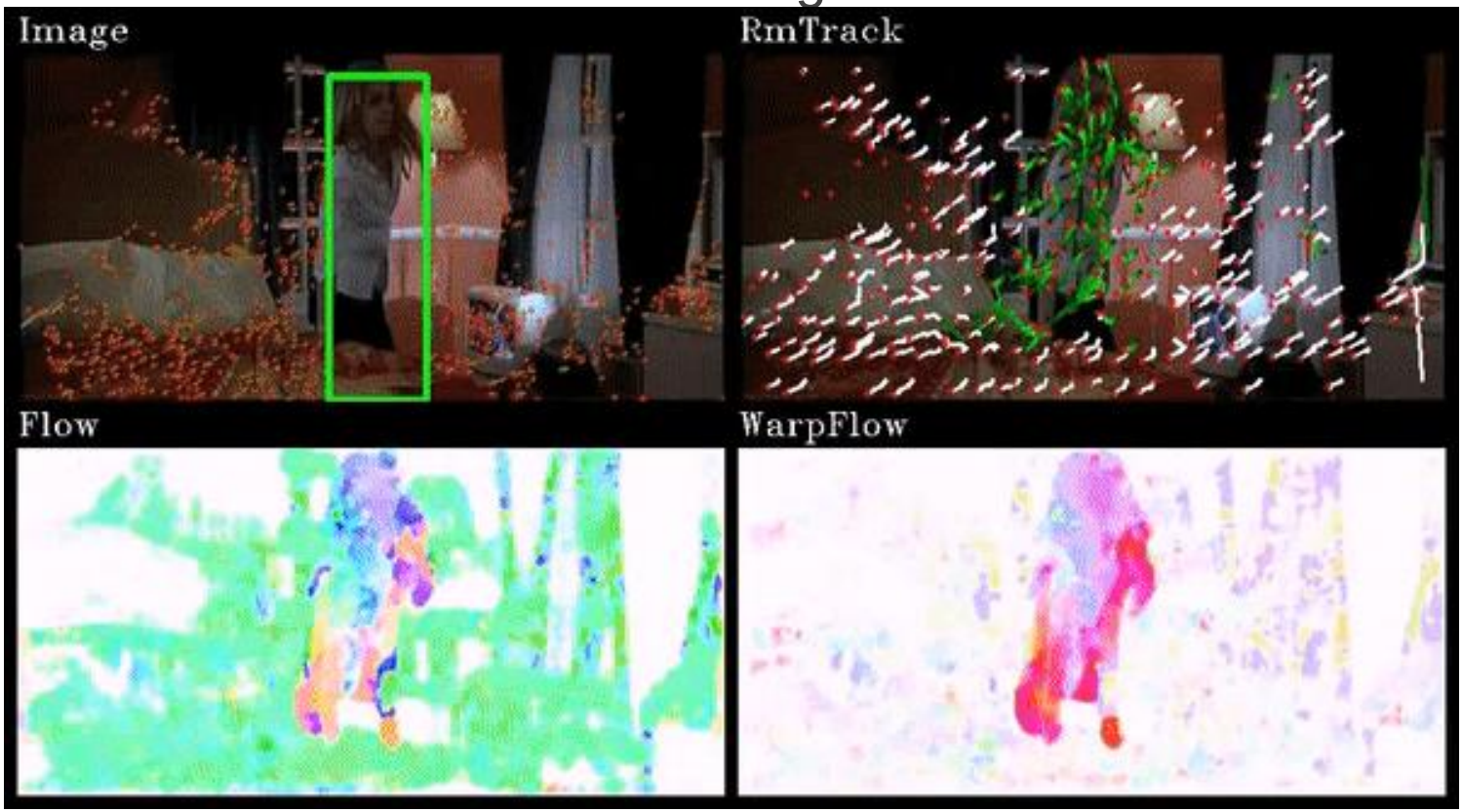
- 3D pose annotation is a much tougher process since specialized equipment and clothing are required.
- Mocap is a common tool use to capture precise 3D skeleton but the challenge in this is often that models created out of such data are hard to be generalized.



Architectures for Action Recognition I

Background

- The earlier years of action recognition (2013) research sees the employment of extracted trajectories and features. These features are then classified using SVM in order to produce a model for action recognition.



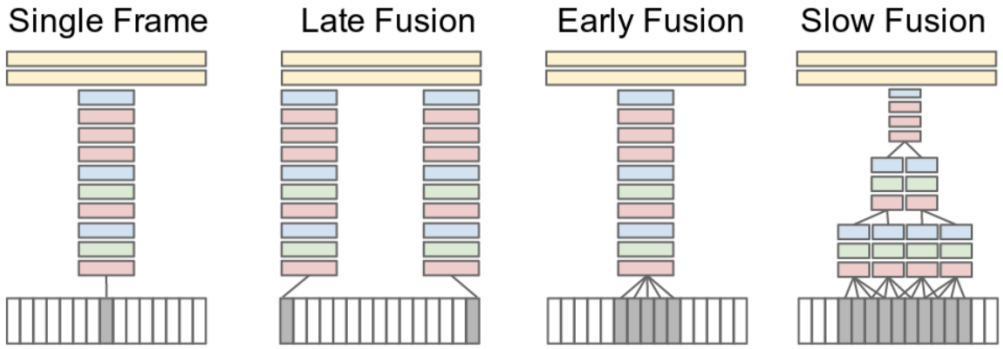
http://lear.inrialpes.fr/~wang/download/demo_iccv13.mp4

Background

- Forward to 2013, two publications were made in 2014 by both Karpathy et. al. and Simonyan and Zisserman that gave birth to a variety of different new architectures:
  - Karpathy proposed the method of performing *Large-scale Video Classification with Convolutional Neural Networks.*
  - Simonyan and Zisserman proposed the use of *Two-Stream Convolutional Networks for Action Recognition in Videos*.

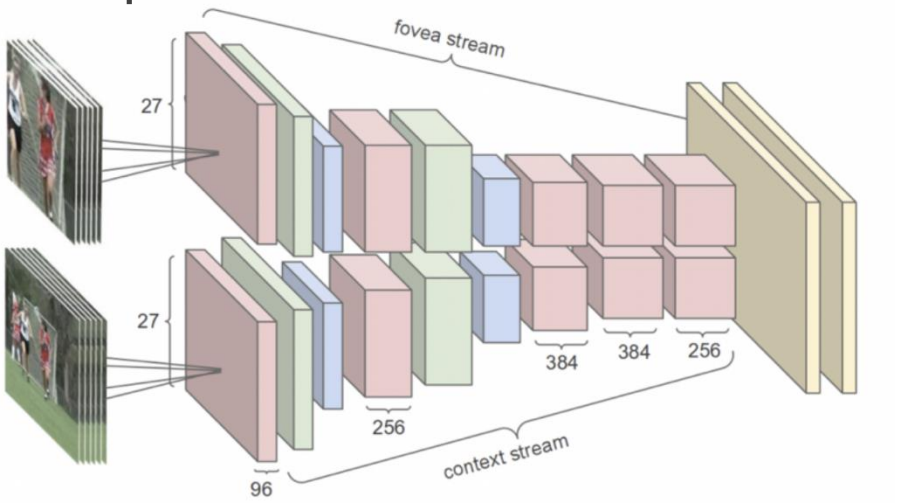
Background

- Karpathy proposed the method of performing *Large-scale Video Classification with Convolutional Neural Networks.*
- Karpathy adopted the use of single-stream approach whereby both spatial and temporal information are retrieved through a sequence of video frames.
- Such an approach allows the use of transfer learning which can dramatically speed up the training process.
- In addition, the method allows for real-time processing of data since the computation intensive optical flow calculation is not employed.



Background

- Simonyan and Zisserman proposed the use of *Two-Stream Convolutional Networks for Action Recognition in Videos*.
- Simonyan and Zisserman adopted the use of a two-stream approach whereby one stream processes a cropped high-resolution stream while the other stream processes a low resolution stream of the fill video before converging into two fully connected layers.
- This approach allowed the author to reduce the number of parameters to be processed.

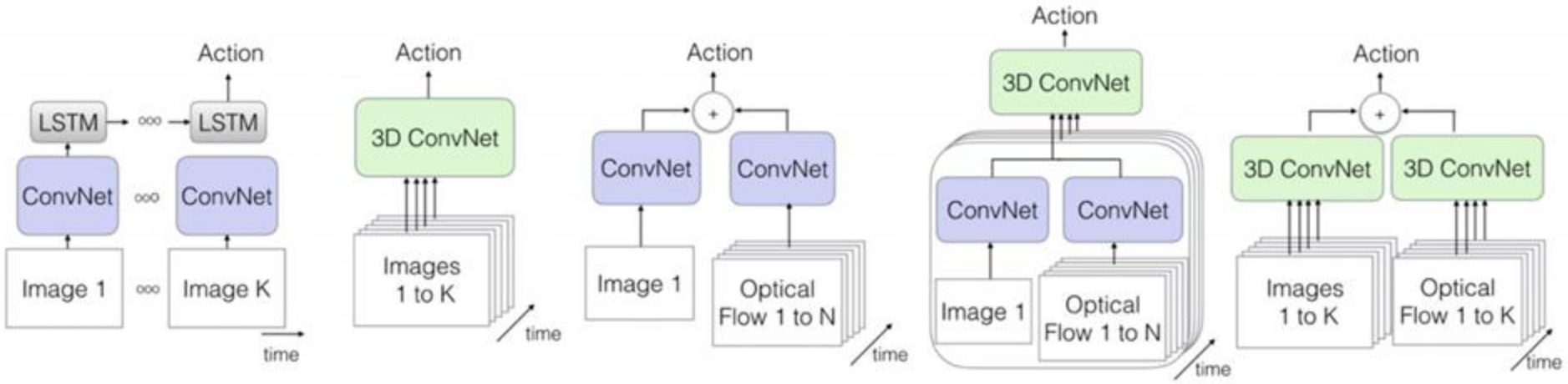


Optical Flow

- Optic flow is about how things are moving in an image at the pixel level.
- Applications of optical flow can range from structures from 3D to image stabilization to our current topic of action detection.
- It is also used to perform segmentation/labelling of images.
- The calculation is performed right down to the individual pixel, on how the change in pixel is happening across the series of frames (commonly known as motion vector).
- One of the challenge faced in the optic world are situations such as a change in lighting condition where an image may be perceived as in motion.
- Another challenge is when objects lack in terms of feature which prevents the analysis from picking up movements, such as a spinning glass ball.

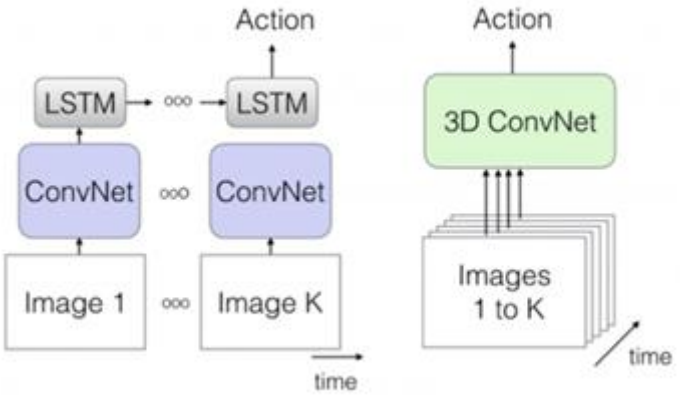
Common Action Recognition Architectures

- There are a few common deep learning architectures that are being used for action recognition and majority of the new architectures built around such architectures.



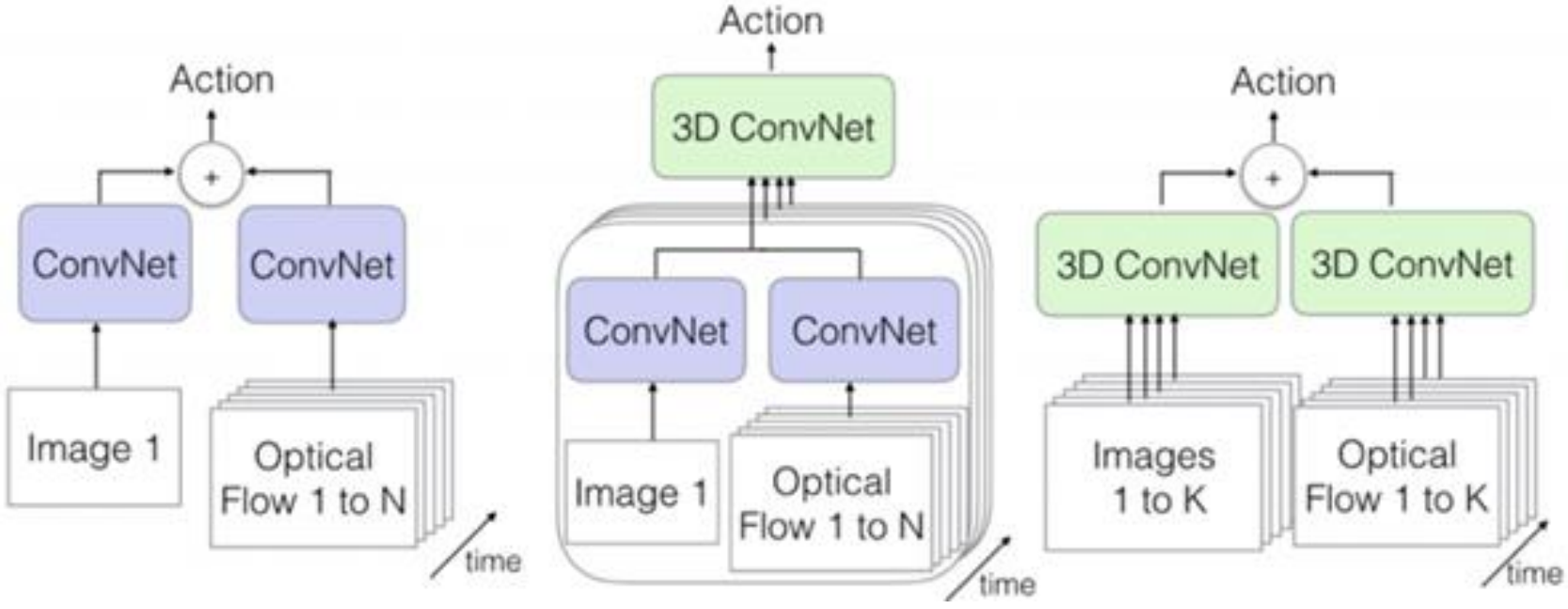
LSTM and 3D ConvNet

- Both the LSTM and 3D ConvNet (CNN) perform direct inference from the series of frames through learning the key features of each action.
- The methods do not make use of optical flow and as a result of this, it is able to allow real-time inference.
- The two architectures are able to learn spatiotemporal features directly through an end-to-end training.
- The end-to-end training refers to training that is performed in a single neural network as compared to a multi-stage approach.



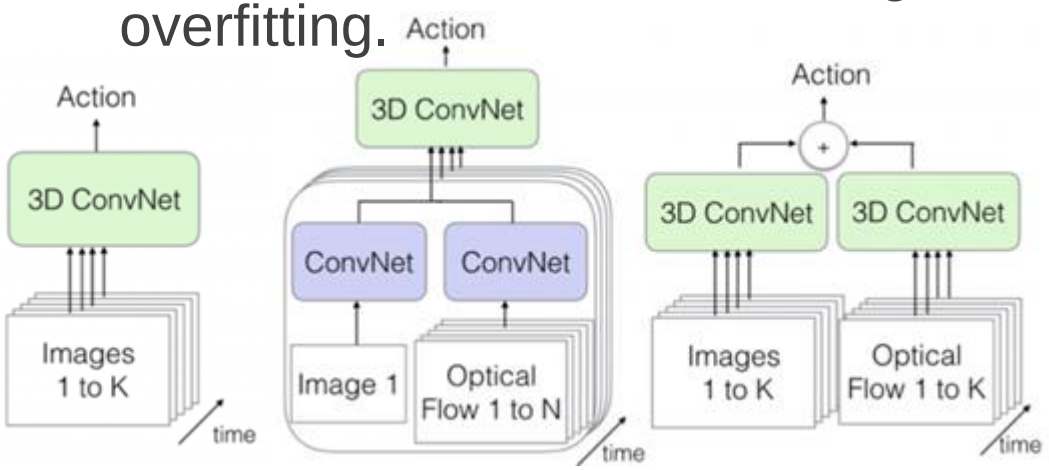
Two-Stream, 3D-Fused Two-Stream, Two-Stream 3D-ConvNet

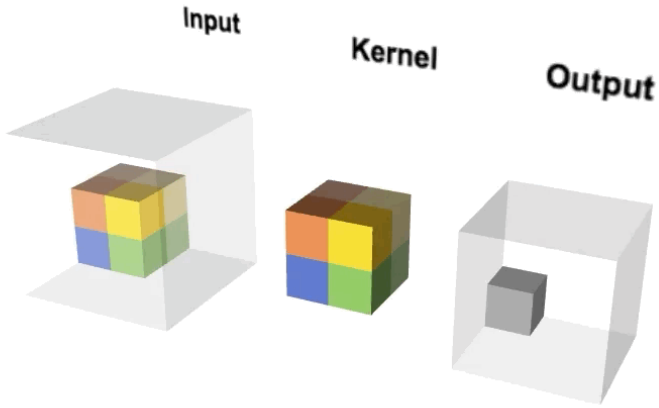
- In contrast to LSTM and 3D ConvNet, the Two-Stream, 3D-Fused Two-Stream and Two-Stream 3D-ConvNet are not able to process videos frames in real-time.
- They require the use of optical flow calculation as part of the process of action inference.
- We will take a quick look at what optical flow is about in the next slide.



3D convolutions

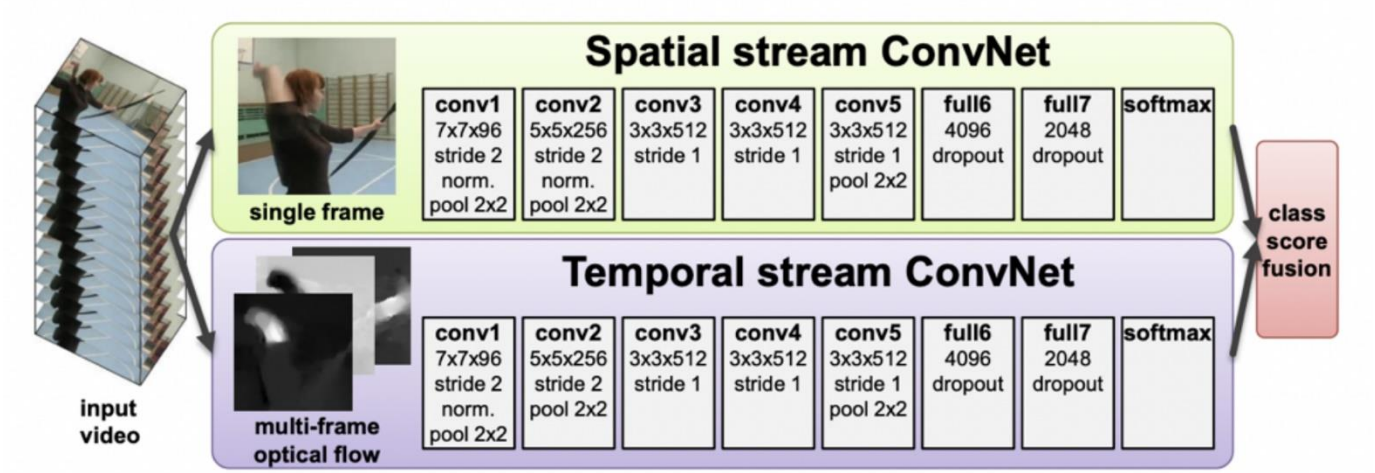
- 3D convolutions are applied on 3D datasets.
- There is a stark contrast of between the amount of datapoint to be processed between a 3D convolution and a 2D convolution.
- The large contrast in datapoints causes a dramatical increase in training time which can often go up to two months at times.
- Similar to a 2D kernel, the 3D kernel also exist in a 3D fashion.
- Another setback of such a large dataset is the problem of overfitting.





Two-Stream Networks

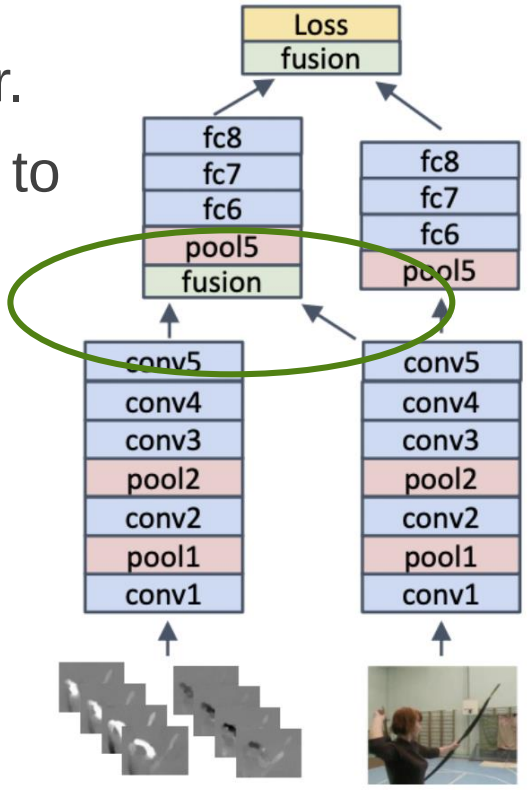
- There is also the concept whereby a Two-Stream approach is being used.
- A Two-Stream approach typically process the same set of data in two parallel streams.
- One stream is commonly used for the purpose of spatial detection while the other stream is used for temporal detection.

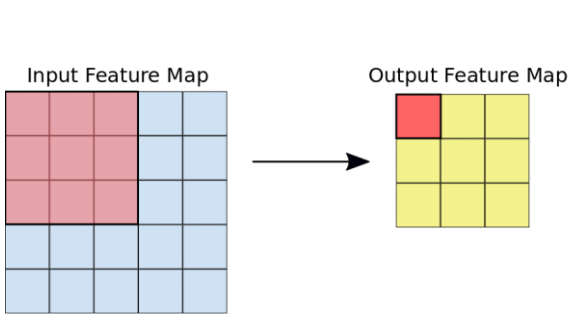


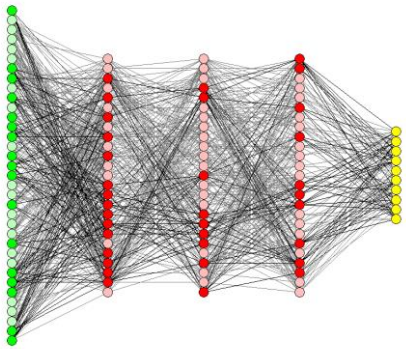
Two-Stream Convolutional Networks for Action Recognition in Videos by Simonyan and Zisserman

Two-Stream Networks

- Another architecture that adopted the Two-Stream networks attempts to merge the two streams just before the fully connected layer.
- The motivation of such a move is to be able to improve the spatial feature differentiator.
- This is done through the combination both spatial and motion feature of an image location.



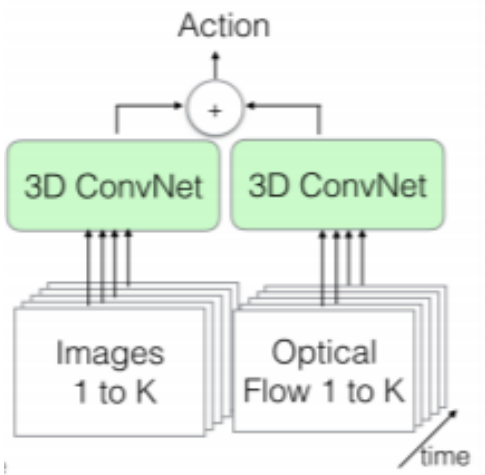




*Convolutional Two-Stream Network Fusion for Video Action Recognition by Zisserman et. al.

Two-Stream Inflated 3D ConvNet (I3D)

- In the recent year, the proposal of inflating a 2D CNN into a 3D CNN was explored.
- The process if inflating 2D CNN into 3D CNN requires the same inflation for the filters, pooling kernels by adding a third dimension of temporal information.
- How the two streams are used is that one stream will process an inflated 3D CNN input while there will be the usual video frames.
- The two networks are trained separated and their predictors averaged.



Lab1
Action Recognition with Two-Stream Inflated 3D ConvNet (I3D)

60 mins Lunch Break

Some interesting videos

https://www.youtube.com/watch?v=bmNaLtC6vkU

https://www.youtube.com/watch?v=Nnf8P5A_saE

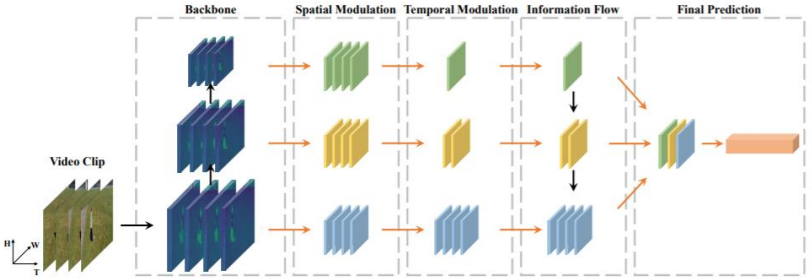
Gentle Reminder:

1. Get your google account ready.
2. Get your google drive and google colab account ready.

Architectures for Action Recognition II

Temporal Pyramid Network

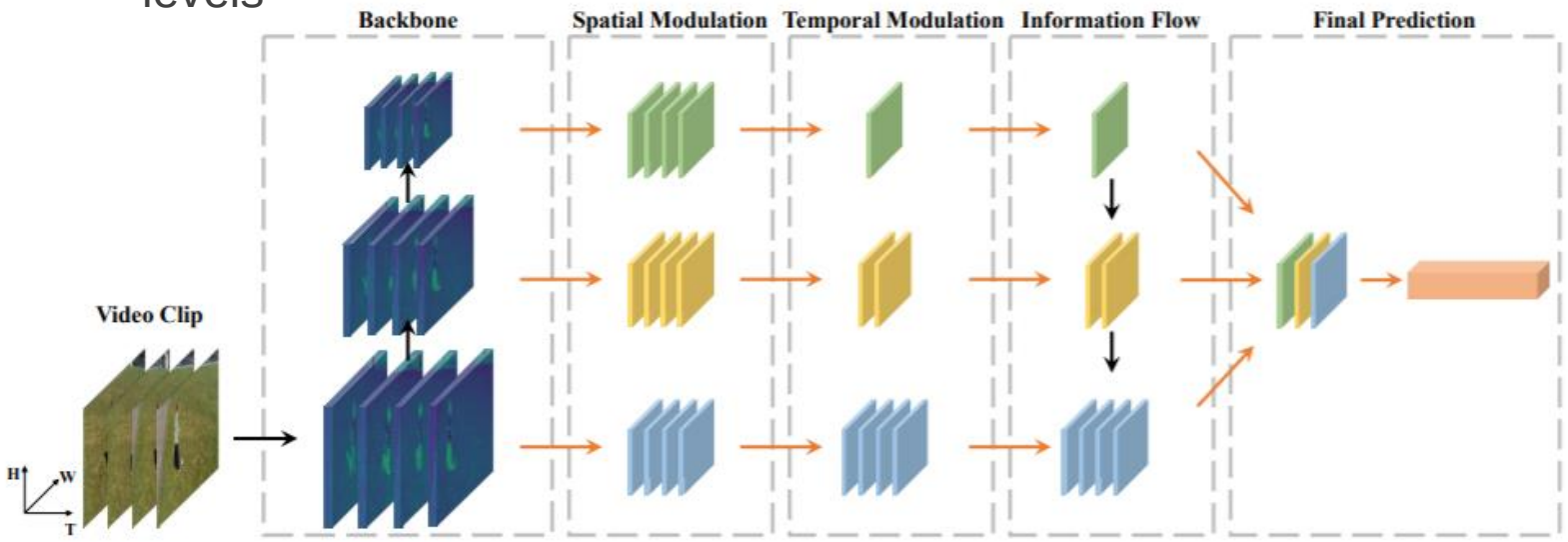
- A modular Temporal Pyramid Network (TPN) was introduced in the recent years in improve the accuracy of action detection.
- This is a versatile module that allows the integration into existing 2D and 3D architectures.
- The temporal pyramid is created through different sampling rates at each pyramid level.
  (Example: 64 frames samples at both 16 and 2 interval thereby producing sets of 4 and 32 frames)
- These frames are then fed into the subnetworks at different levels and the output from these levels are then combined at the end for the final prediction.
- The motivation behind this is to produce a generalized model that is capable of action inference on different temporal scales.



Temporal Pyramid Network for Action Recognition – Ceyuan Yang et. al.

Temporal Pyramid Network

- The framework of TPN is broken down as such:
  - Backbone network will extract multiple level of features.
  - Spatial modulation performs scaling and alignment of the spatial features.
  - Temporal modulation down samples the features (reduces the frames)
  - Information flow aggregates the features
  - Final Prediction rescales and concatenates the product from all levels



Temporal Pyramid Network for Action Recognition – Ceyuan Yang et. al.

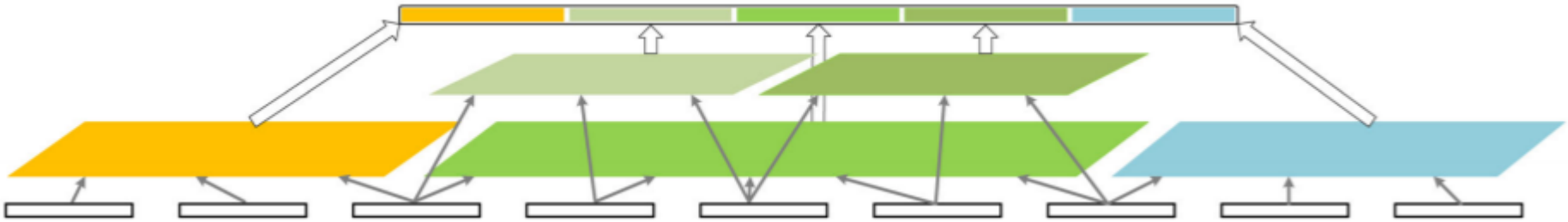
Structured Segment Networks

- In a related architecture, a Structured Segment Network (SSN) was being proposed for action recognition.
- The SSN first employs a proposal method to create a set of proposals with varying number of frames.
- Each proposal is said to contain the start, course and end of an action.
- The Structured Temporal Pyramid Pooling (STPP) is then applied to each proposal (much similar to TPN).

Temporal Action Detection with Structured Segment Networks - Yue Zhao et. Al.

Structured Segment Networks

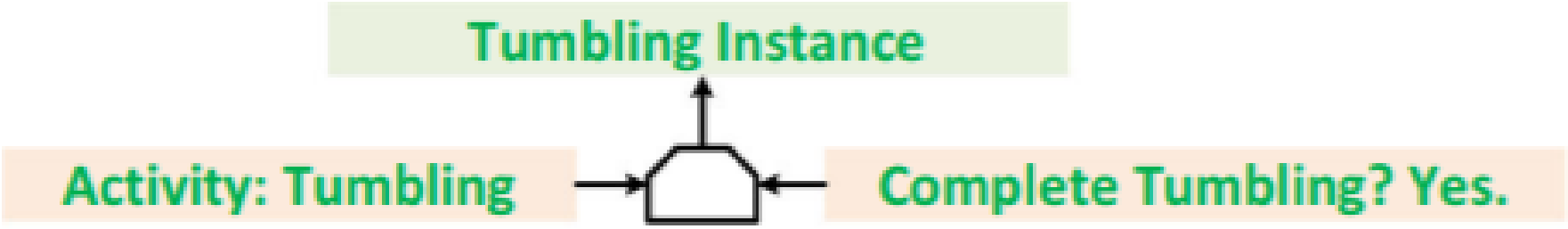
- The structured temporal pyramid pooling is performed as follows (ref TPN):
  1. splitting the proposal into three stages
  2. building temporal pyramidal representation for each stage
  3. building global representation for the whole proposal by concatenating stage-level representations



Temporal Action Detection with Structured Segment Networks - Yue Zhao et. Al.

Structured Segment Networks

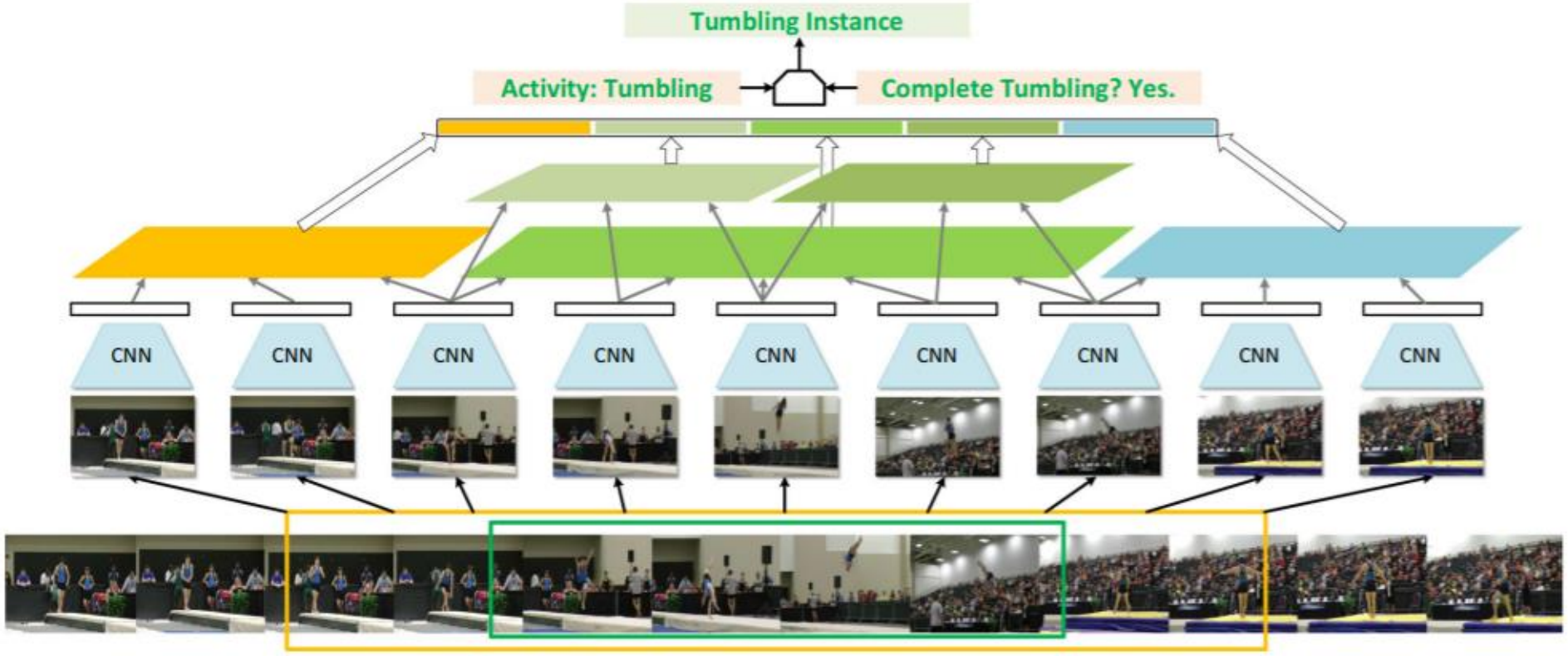
- The final part required two classifiers to be applied to the output from STPP.
- The first is the activity classifier and the second is the completeness classifier.
- This will then produce one of the action recognized from the video.
- Other proposals that are determined as background or incomplete are filtered away.
- The whole process is integrate and trained as an end-to-end system.



Temporal Action Detection with Structured Segment Networks - Yue Zhao et. Al.

Structured Segment Networks

- The complete flow of the SSN.



Temporal Action Detection with Structured Segment Networks - Yue Zhao et. Al.

Lab2
Action Recognition with MMAction2

https://youtu.be/ooqiY7hFWE8

15 Mins Break

https://youtu.be/_sBBaNYex3E

Which one of the following would strike you as an important events?

1. Two man running
2. One man holding up a machete (machete a.k.a. parang)
3. One man holding up a sword
4. One man running

Lab3: Creating your own action recognition model

Survey Time!

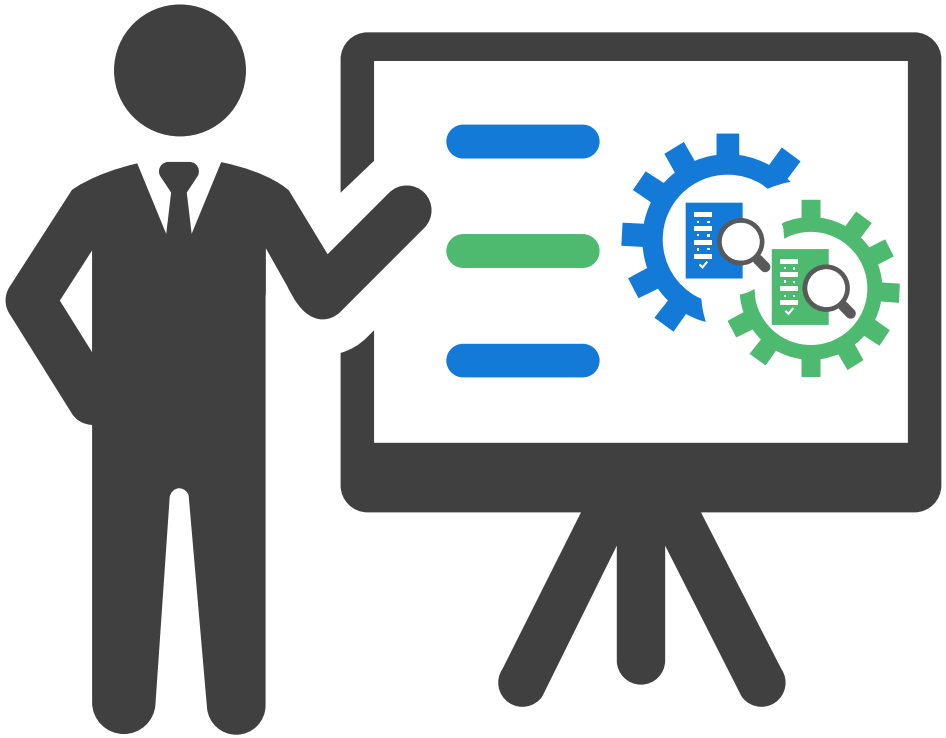
Please help to complete the survey before you leave the session. ^_^

(TRAQOM)

References

- https://machinelearningmastery.com/how-to-develop-rnn-models-for-human-activity-recognition-time-series-classification/
- https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21
- https://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9
- https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/
- https://towardsdatascience.com/deep-learning-architectures-for-action-recognition-83e5061ddf90
- https://github.com/gtoderici/sports-1m-dataset/blob/wiki/ProjectHome.md
- https://medium.com/agile-lab-engineering/going-deep-into-real-time-human-action-recognition-a99483b74ded

Summary



Email
Zack_toh@rp.edu.sg

Telegram
@zacktohsh

Source code:

Thank you!