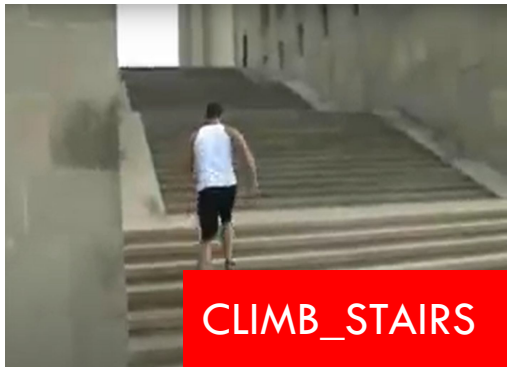
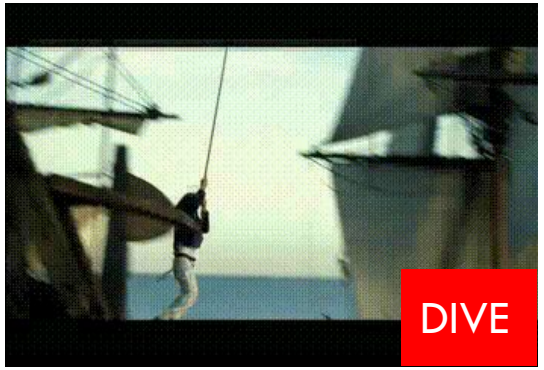
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

HUMAN ACTIVITY RECOGNITION

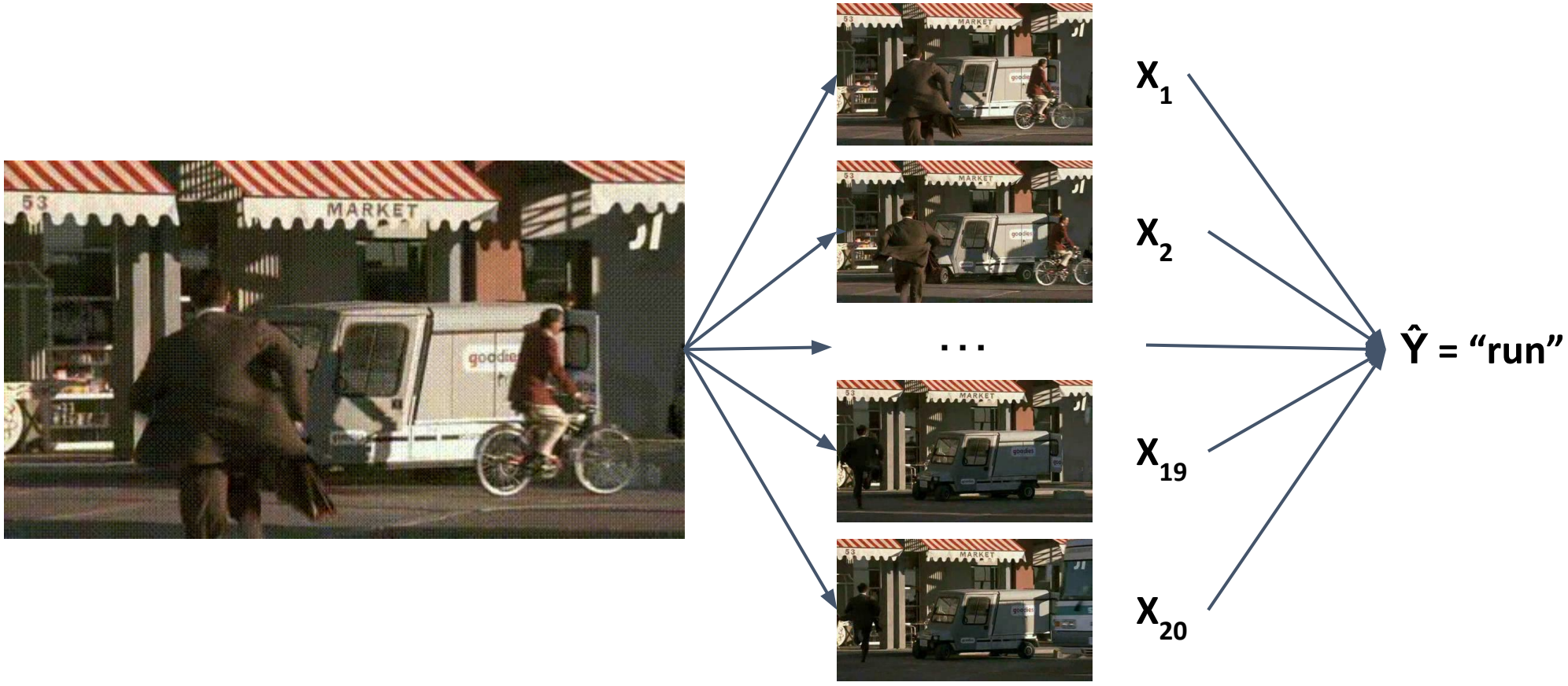
Team: Giannelli Alessio, Imbonati Lorenzo, Valoti Davide





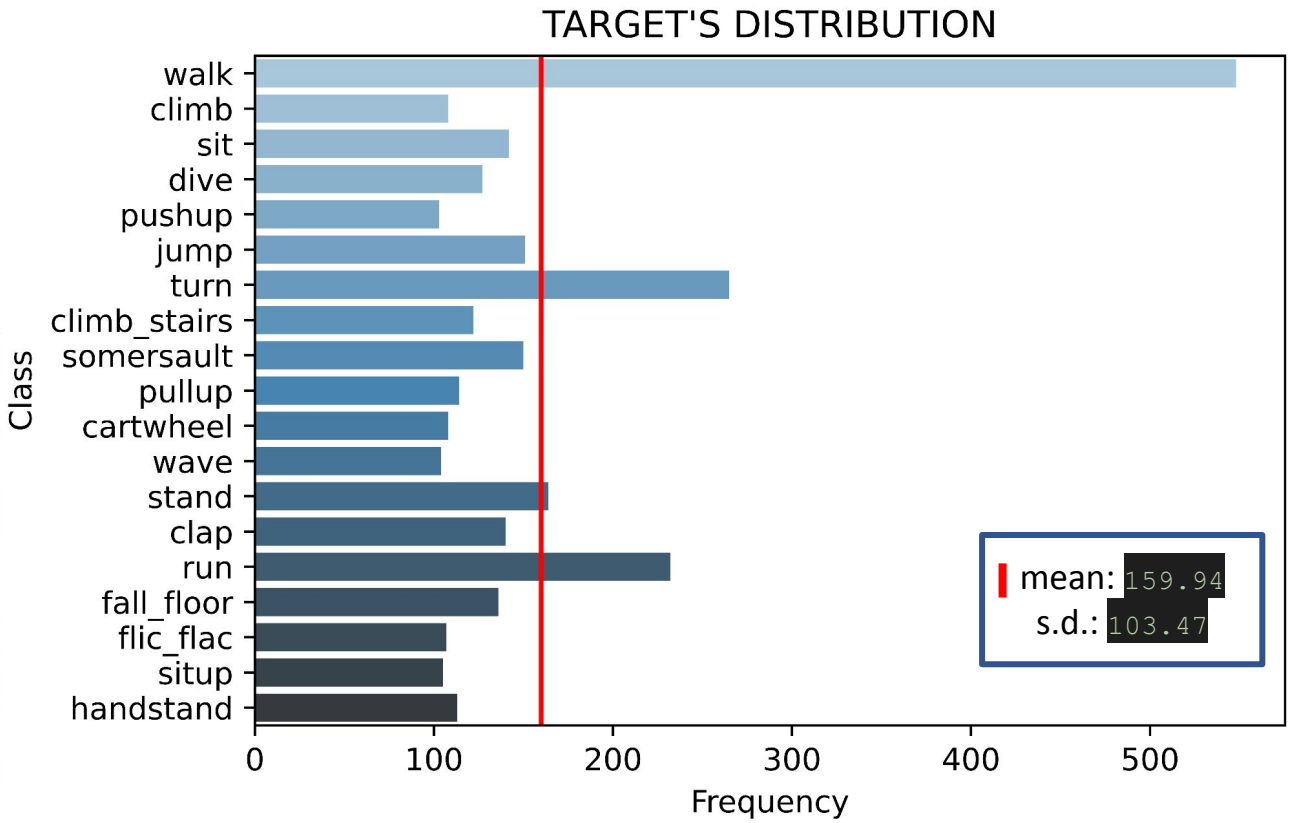
PROJECT GOAL

The goal of the project is to develop a classification algorithm that exploits the main deep learning techniques in order to predict and recognize the simplest human actions.



DATASET

The selected dataset is named 'HMDB - Human Emotion DB'. Each observation corresponds to one video, for a total of 6849. Each video has associated one of 51 possible classes, each of which identifies a specific human behavior.
Due to computational problems we have chosen only 19 classes on which to train the human activity recognition algorithm.



INPUT & PREPROCESSING

SIZE CONSTANTS
Image Height = 64
Image Width = 96
N° Frames
per video = 20
N° of classes = 19

Frames_Extraction()

Resized Frame
[64 x 96]
Normalized Frame

Create_Dataset()

Features, i.e. normalized frames
Labels, i.e. tags of category

np.array() +
to_categorical()

Features shape (3038, 20, 96, 64, 3)
Labels' Matrix (3038, 19) - [one hot encode]

train_test_split()

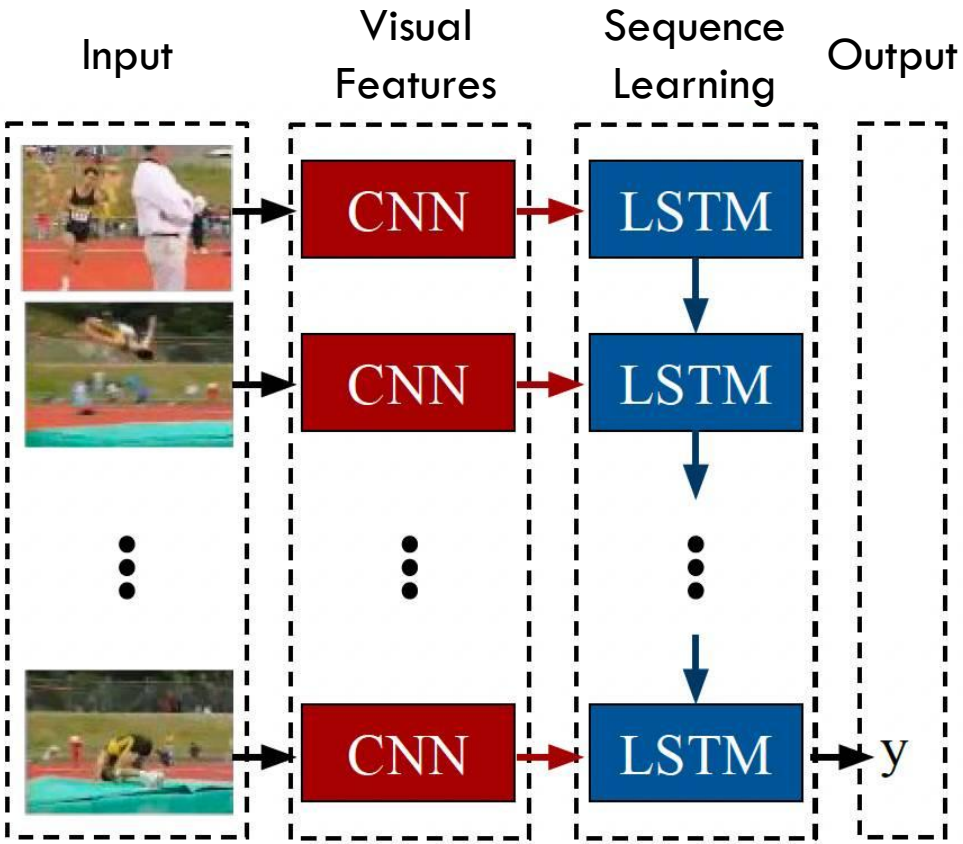
Features Train (2278, 20, 96,64,3)
Features Test (760, 20, 96, 64, 3)

Labels Train (2278, 19)
Labels Test (760, 19)

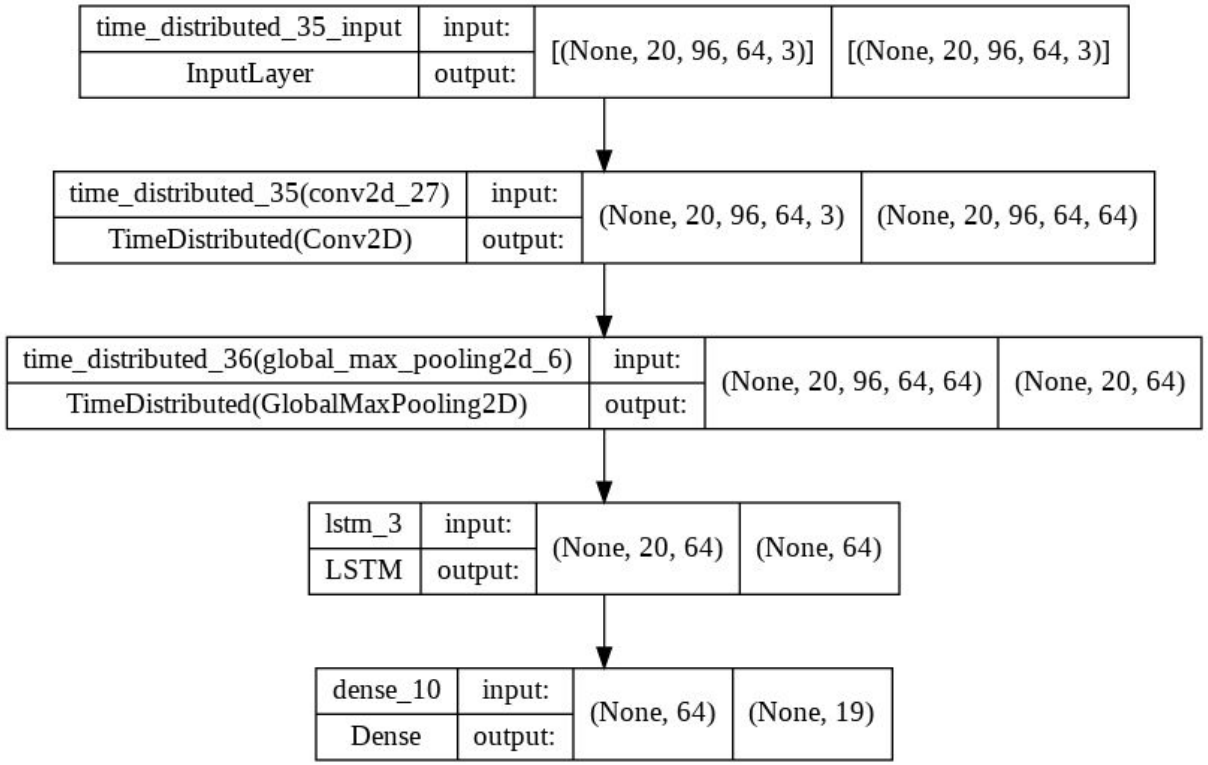
Parameters: - stratify;
- shuffle;
- size = 0.25

LRCN APPROACH

Def LRCN: a class of architectures which combines Convolutional layers and Long Short-Term Memory (LSTM).



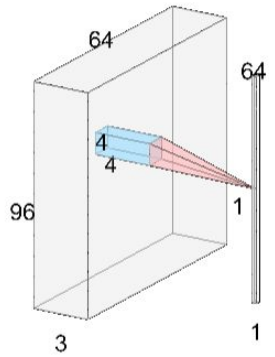
IMPLEMENTED MODELS

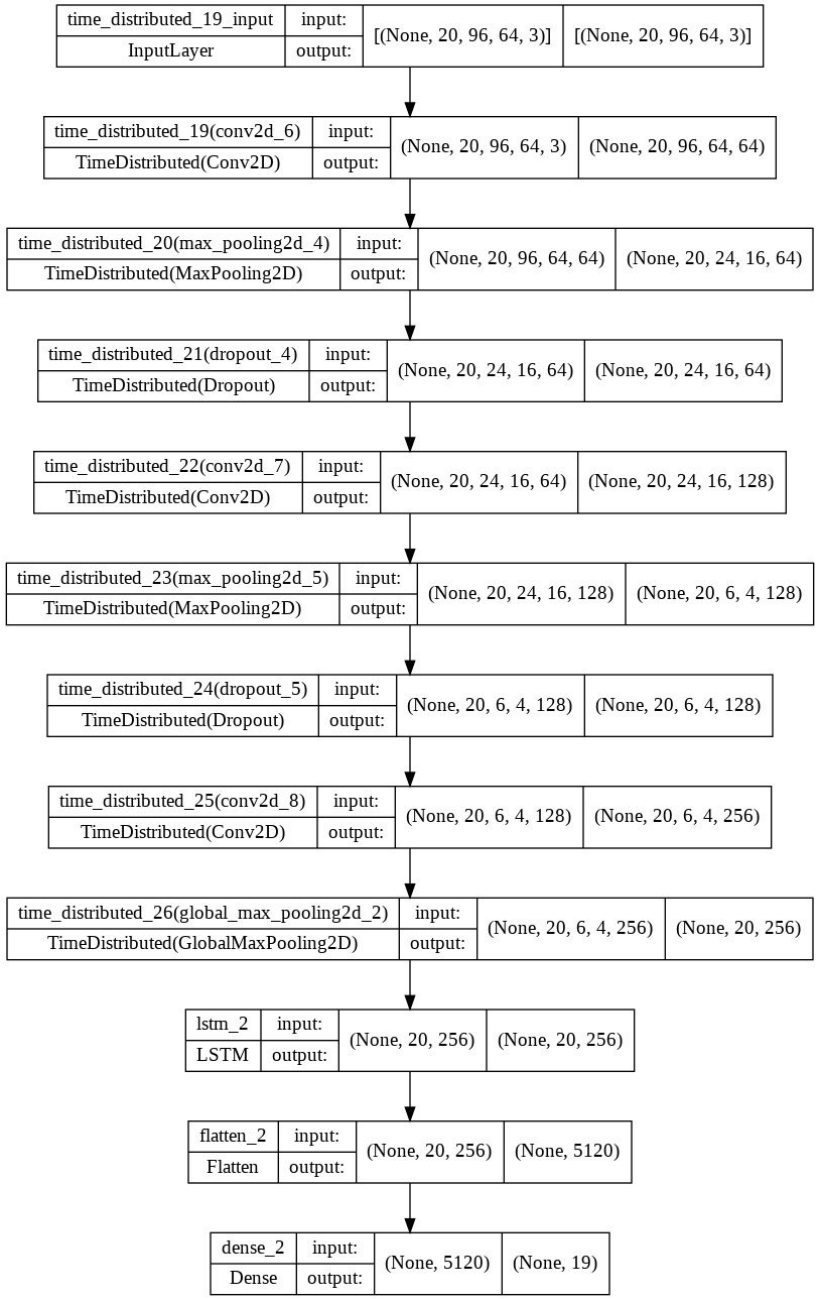


BASIC LRCN

Basic Structure:

1. Convolutional2D Layer
2. LSTM Layer
3. Dense Layer [fully - connected]

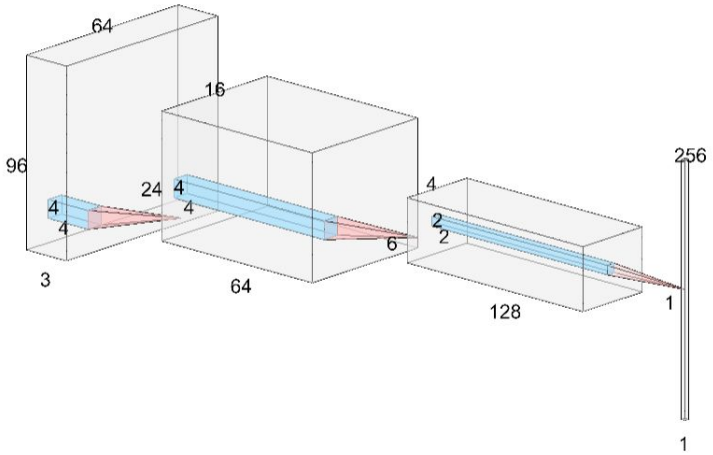




ADVANCED LRCN

Basic Structure:

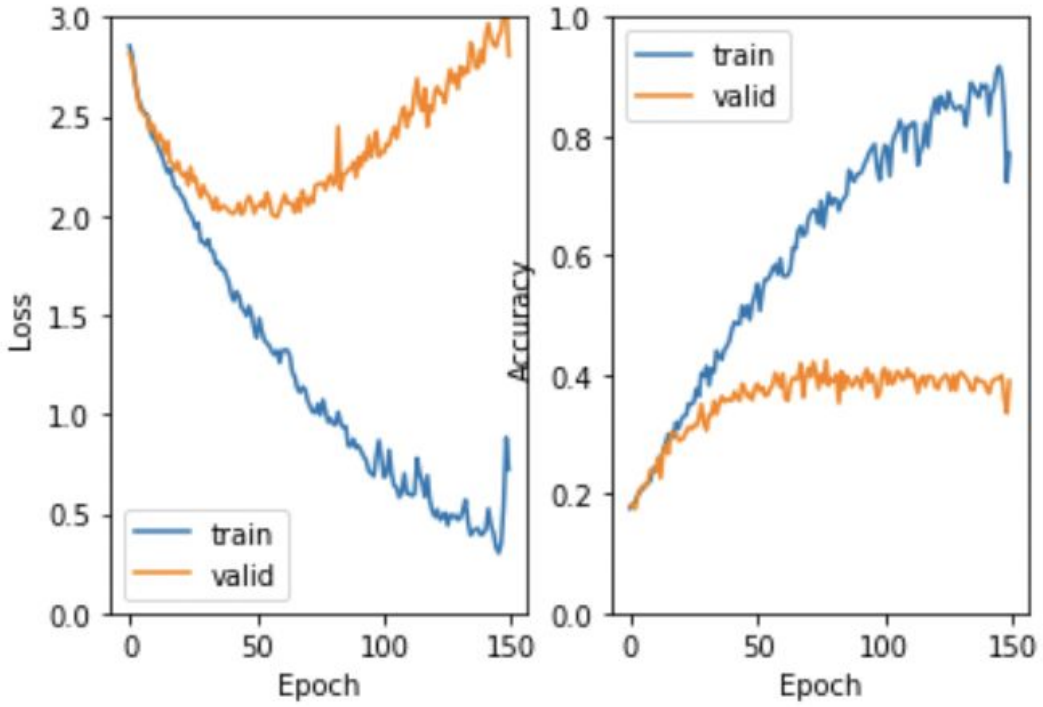
1. Three Convolutional2D Layers
2. LSTM Layer
3. Dense Layer [fully - connected]



RESULTS

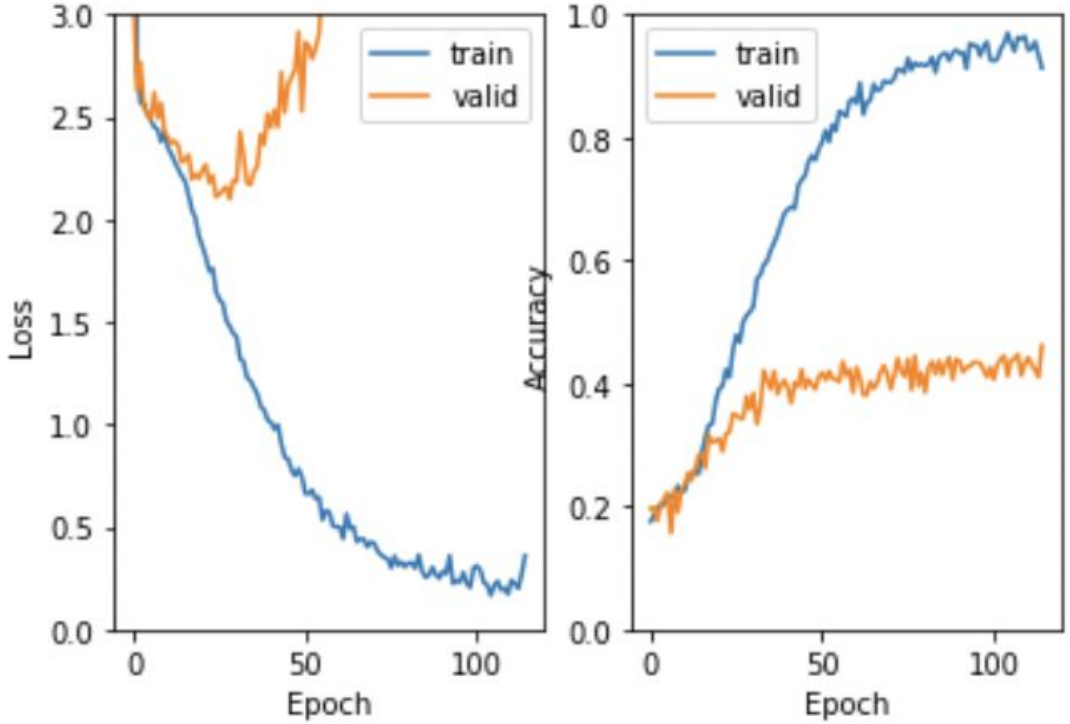
BASIC LRCN

loss: 0.7570 - accuracy: 0.7673 - val_loss: 2.7401 - val_accuracy: 0.3447

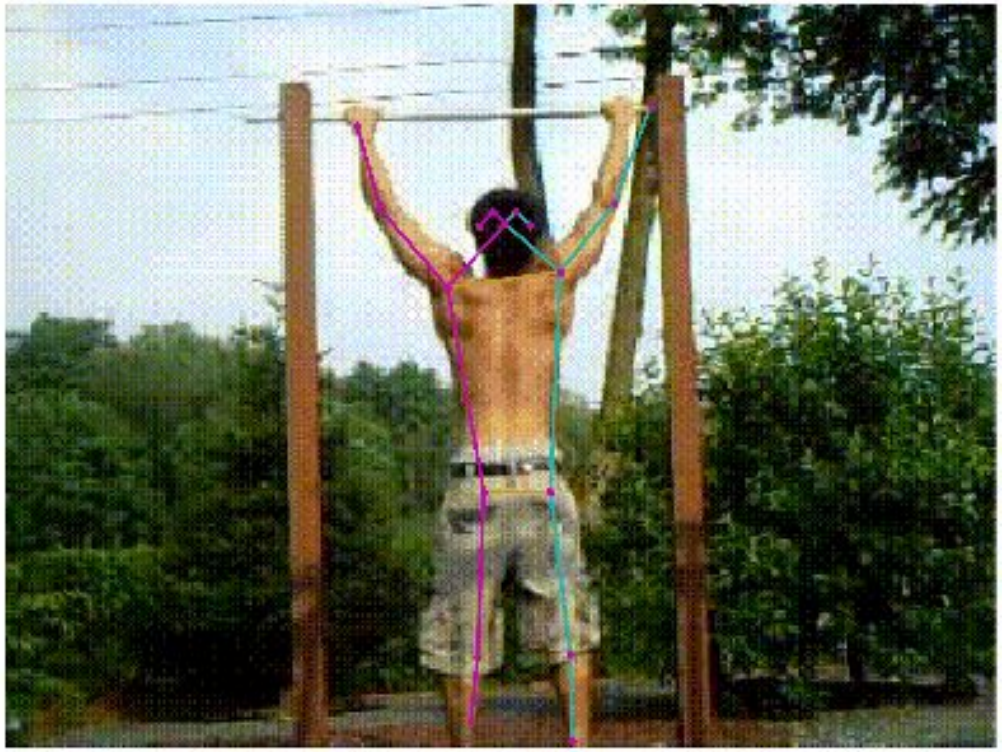


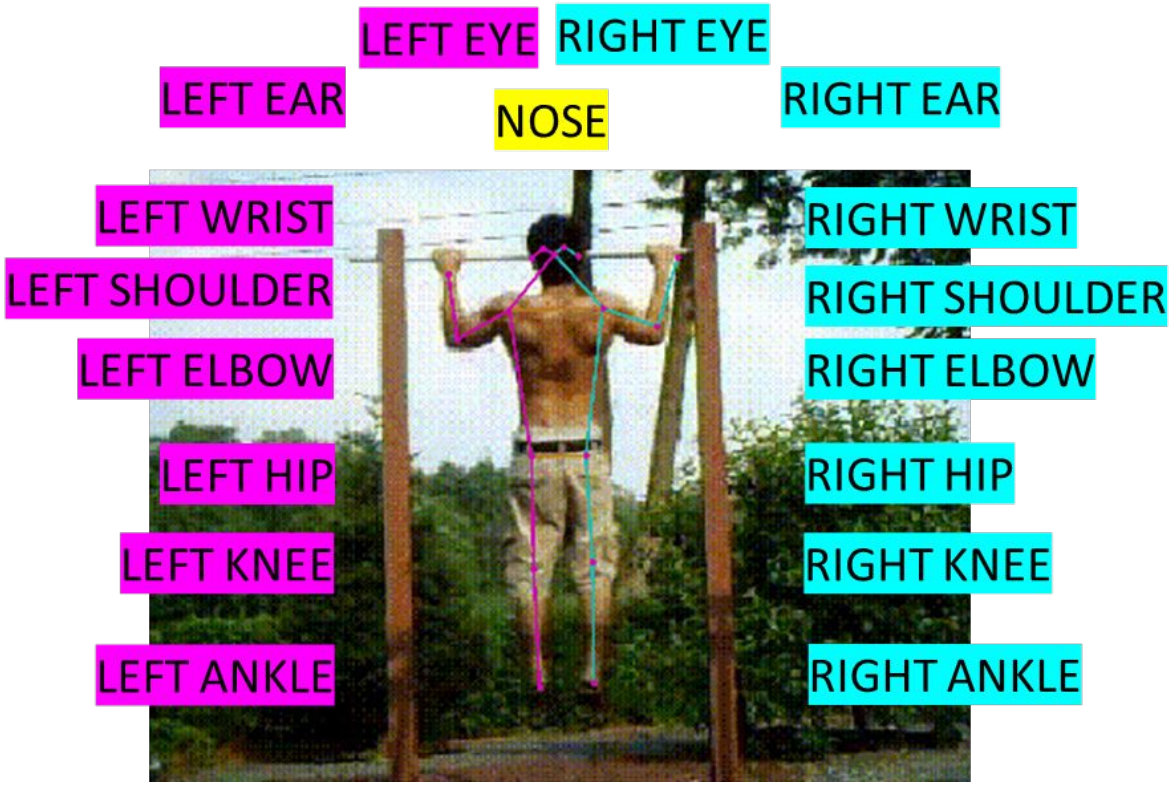
ADVANCED LRCN

loss: 0.1658 - accuracy: 0.9701 - val_loss: 4.2734 - val_accuracy: 0.4118

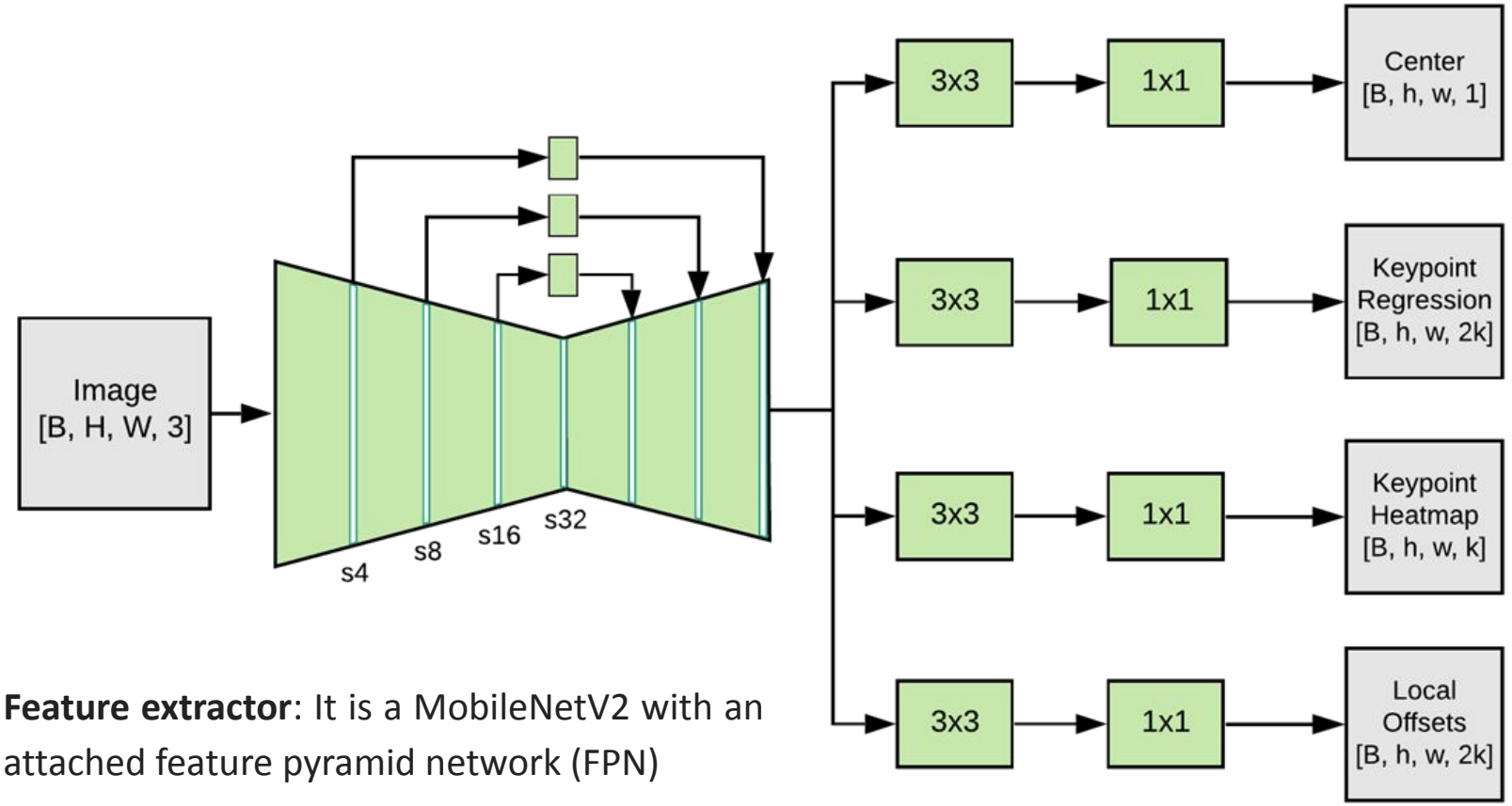


MOVENET APPROACH





INPUT

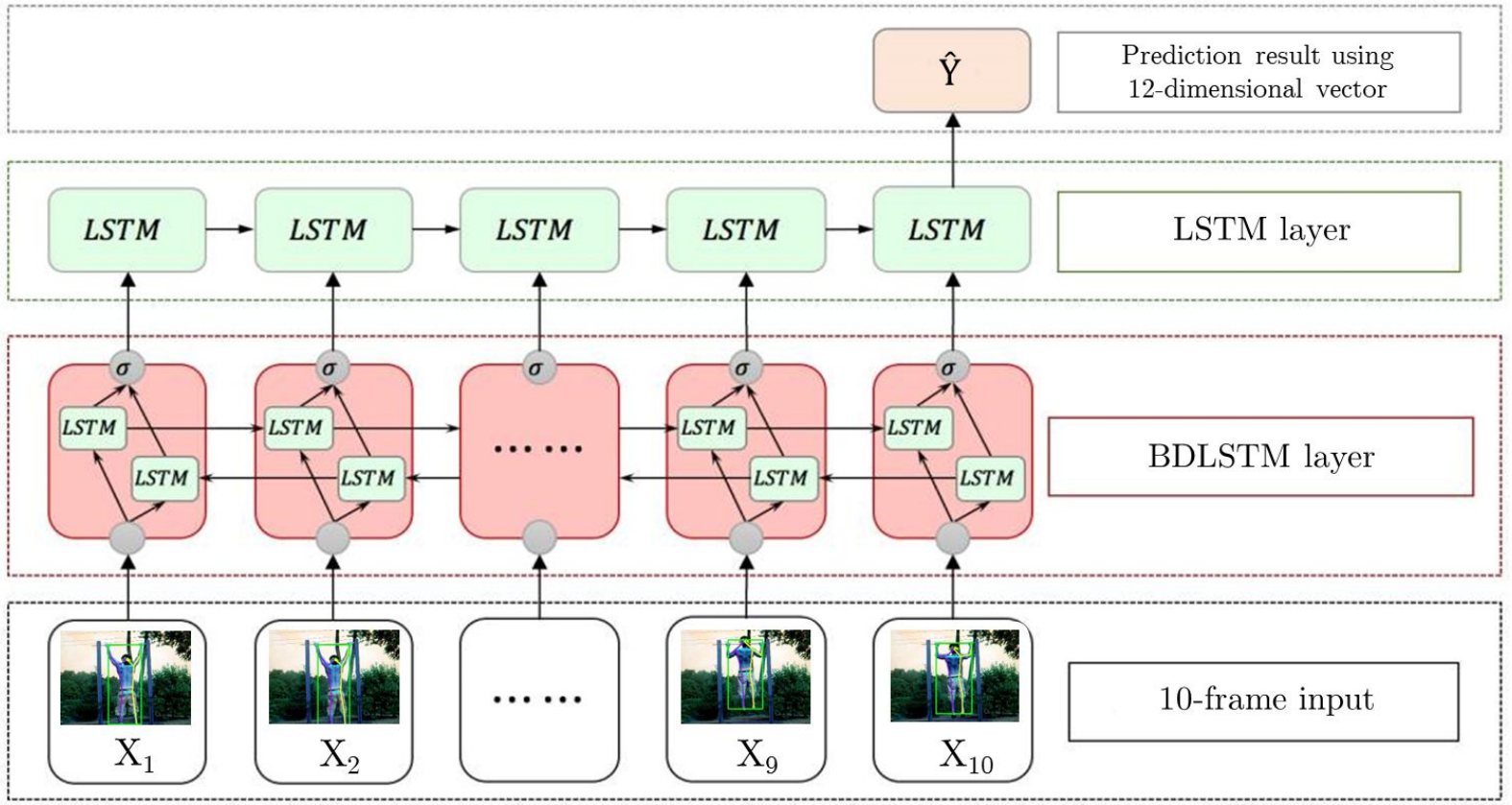


Feature extractor: It is a MobileNetV2 with an attached feature pyramid network (FPN)

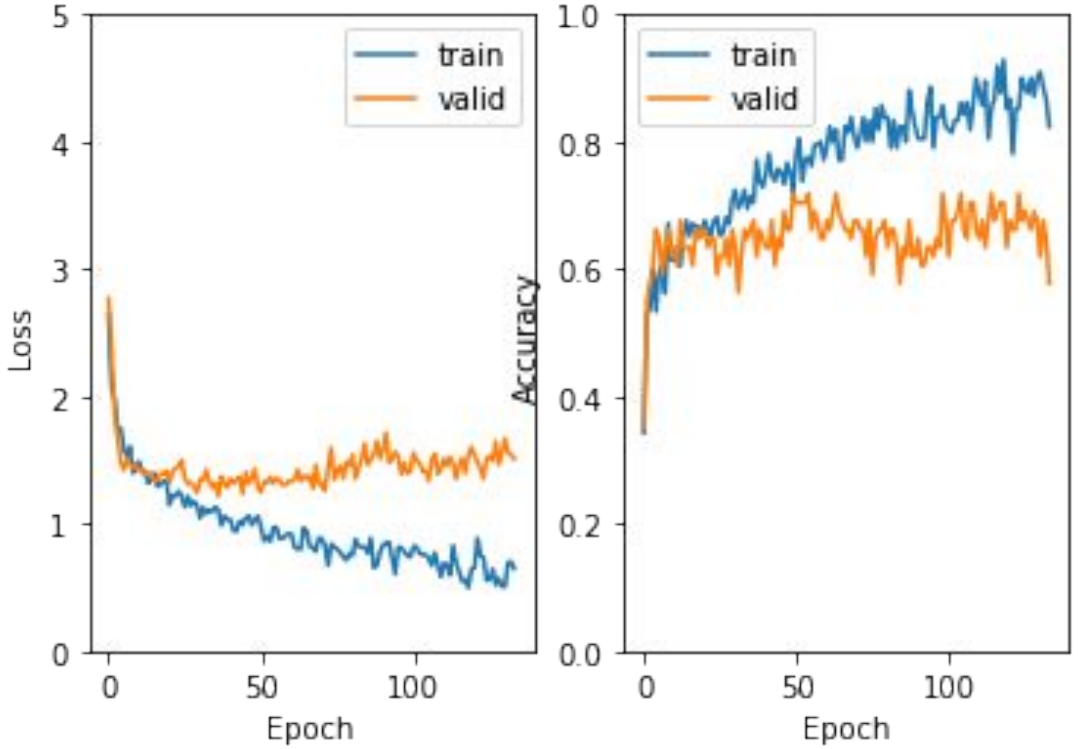
Set of prediction heads: Four prediction heads attached to the feature extractor

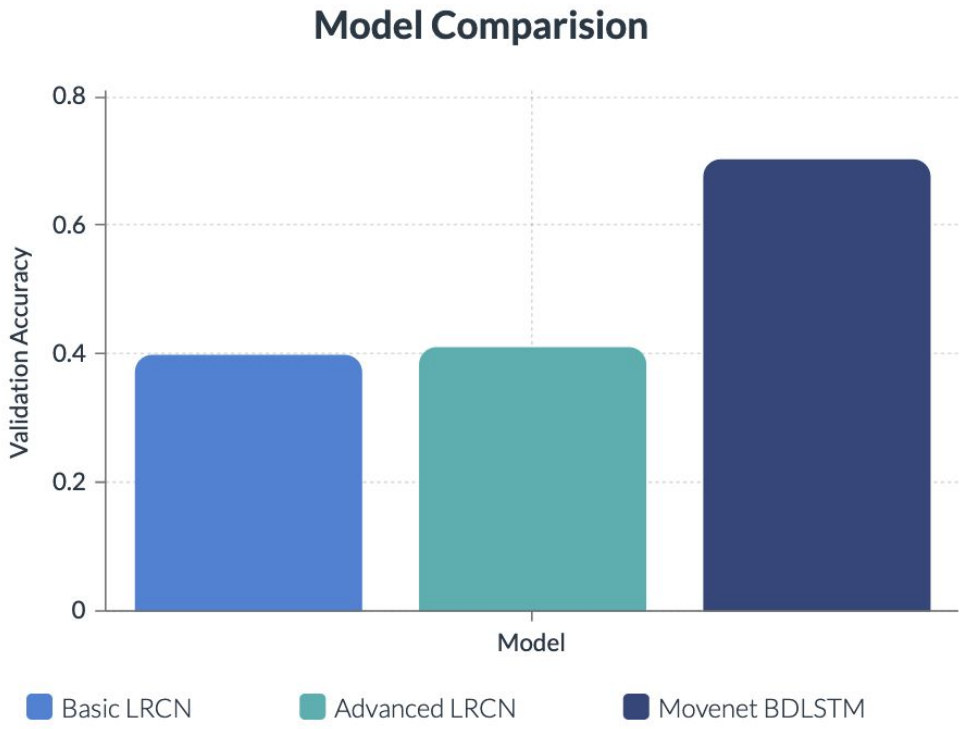
BDLSTM ARCHITECTURE

Def DBLSTM: The idea of BDLSTMs comes from bidirectional RNN, which processes sequence data in both forward and backward directions with two separate hidden layers. BDLSTMs connect the two hidden layers to the same output layer.



RESULTS


train
valid
Loss
Epoch
train
valid
Accuracy
Epoch


Model Comparision
Validation Accuracy
Model
Basic LRCN
Advanced LRCN
Movenet BDLSTM

CONCLUSIONS

To sum up, we have seen that the best approach for human activity recognition videos is the last one, where Movenet's output is used as input for a BiDirectional LSTM. This method allow us to reach a good level of validation accuracy, over 70%.

CRITICALITIES

- Low quantity and quality of video used as input
- Computational power
- Underperformance of CNNs models

FUTURE DEVELOPMENTS

- Development of a MoveNet model able to handle cropped human bodies
- Train of the best model to other categories of activities, like the ones with objects

REFERENCES

[1] Deep Learning Models for Human Activity Recognition

https://machinelearningmastery.com/deep-learning-models-for-human-activity-recognition/

[2] Long-term Recurrent Convolutional Networks for Visual Recognition and Description

https://arxiv.org/abs/1411.4389?source=post_page1

[3] Long-term Recurrent Convolutional Network for Video Regression

https://towardsdatascience.com/long-term-recurrent-convolutional-network-for-video-regression-12138f8b4713

[4] Long-term Recurrent Convolutional Networks

https://jeffdonahue.com/lrcn/

[5] Next-Generation Pose Detection with MoveNet

https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html

THANK YOU
FOR YOUR ATTENTION