Action Recognition in Videos using Deep Learning

Deep Learning course

MSc in AI

Georgios Batsis

About project

Action Recognition:

- Classifying the activity being performed by a human
- We need a set of evidence to recognize an action → Video classification

Video:

- A signal which combined spatial and temporal information
- Sequence of images-frames

MODELS

Baseline

Step 1 → Feature Extraction of each frame: Texture information → GLCM features:

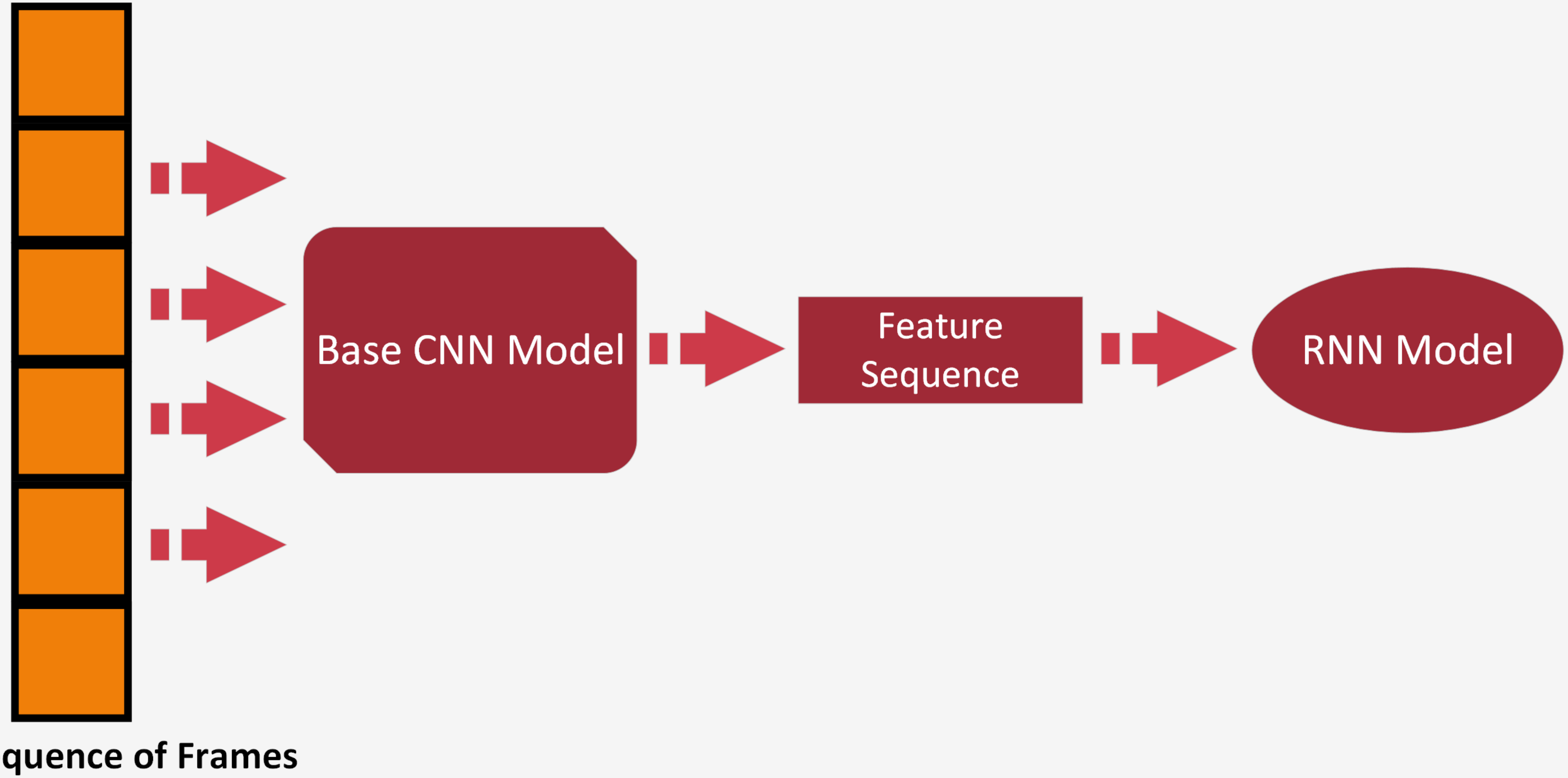
- Contrast
- Dissimilarity
- Homogeneity
- ASM
- Energy
- Correlation

Step 2 → Temporal Aggregation: Statistics → Mean & Std

Step 3 → Definition of a classifier pipeline:

- Standard Scaler
- SVM with RBF Kernel

Going Deeper...



Experiments: Step 1 → Choose Base CNN Model

Most popular pretrained models:

- VGG
- ResNet
- MobileNet.

Choose which layers will be left frozen:

CNNs consist of:

- Convolutional Block → Convolutional and Pooling Layers
- Classifier → Fully Connected Layers

Freezing options:

- Last (or other) Convolutional Layer
- Penultimate Layer of classifier

Experiments: Step 2 → Deep Model Construction

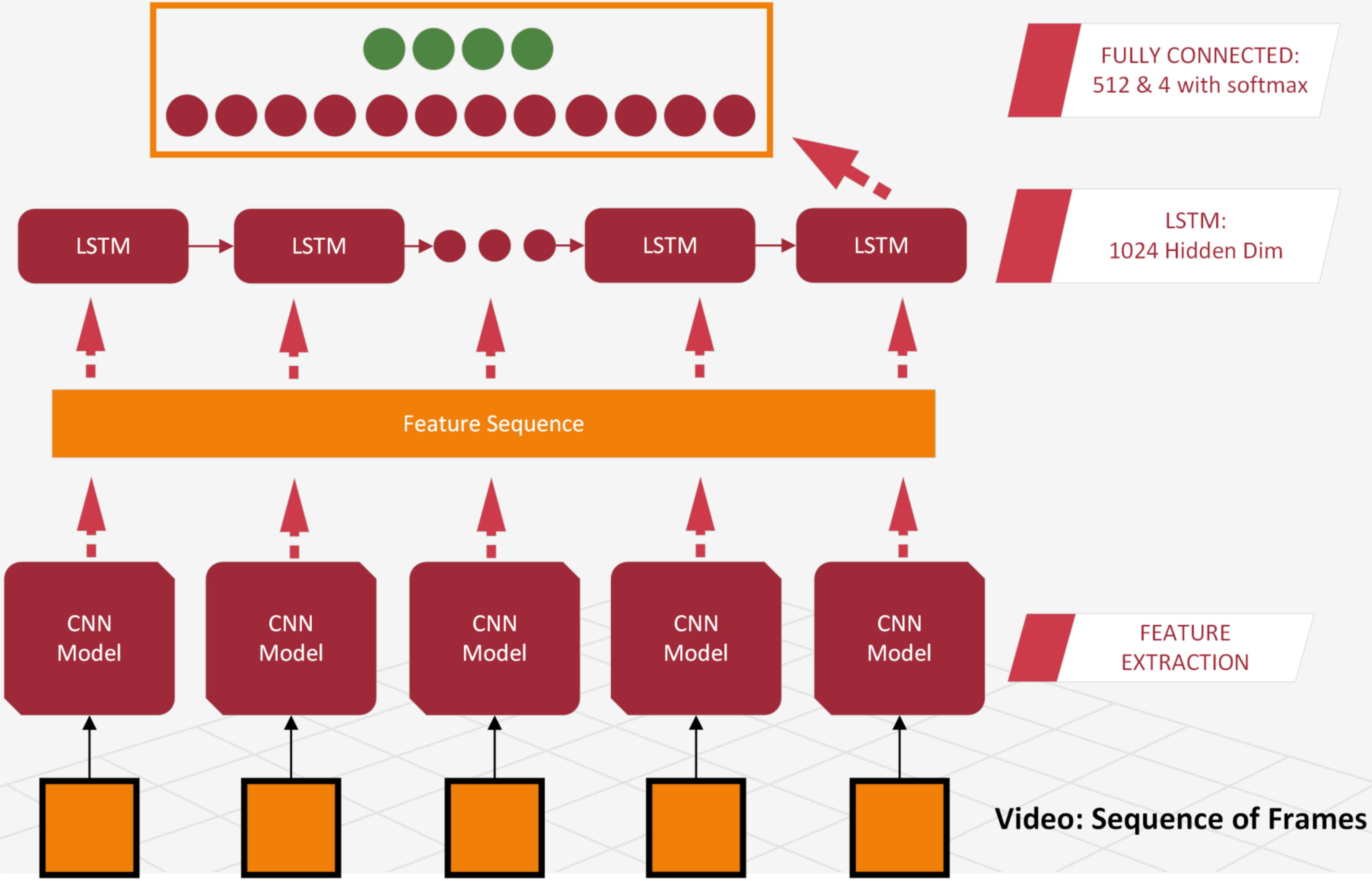
After pre-trained CNN selection, decide if:

- Use the extracted features directly
- Add trainable layer(s) after feature extraction

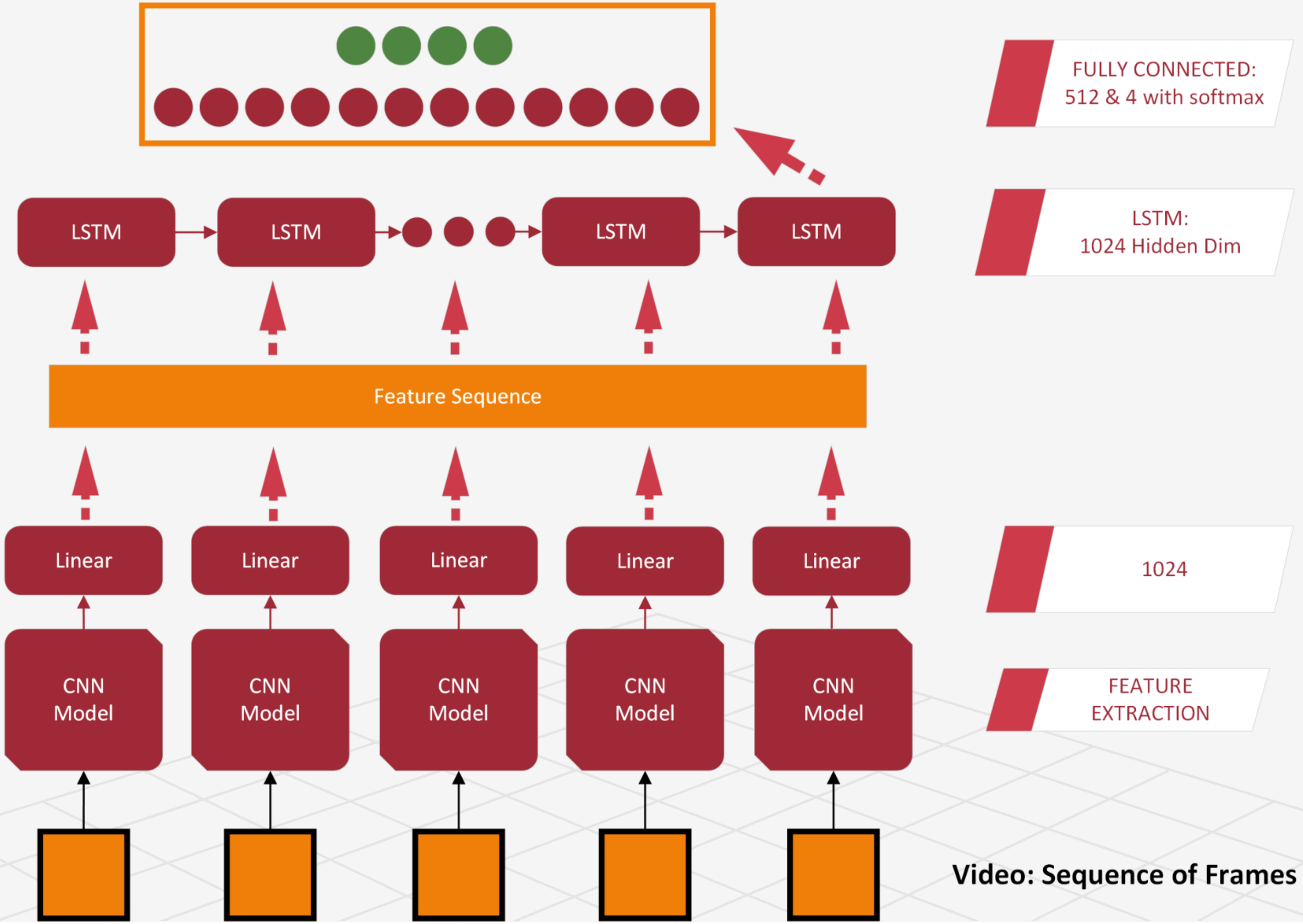
Define the RNN-based part:

- Model (e.g. LSTM)
  - Hidden state dimension
  - Number of layers
  - Direction
- Final classifier

Approach A

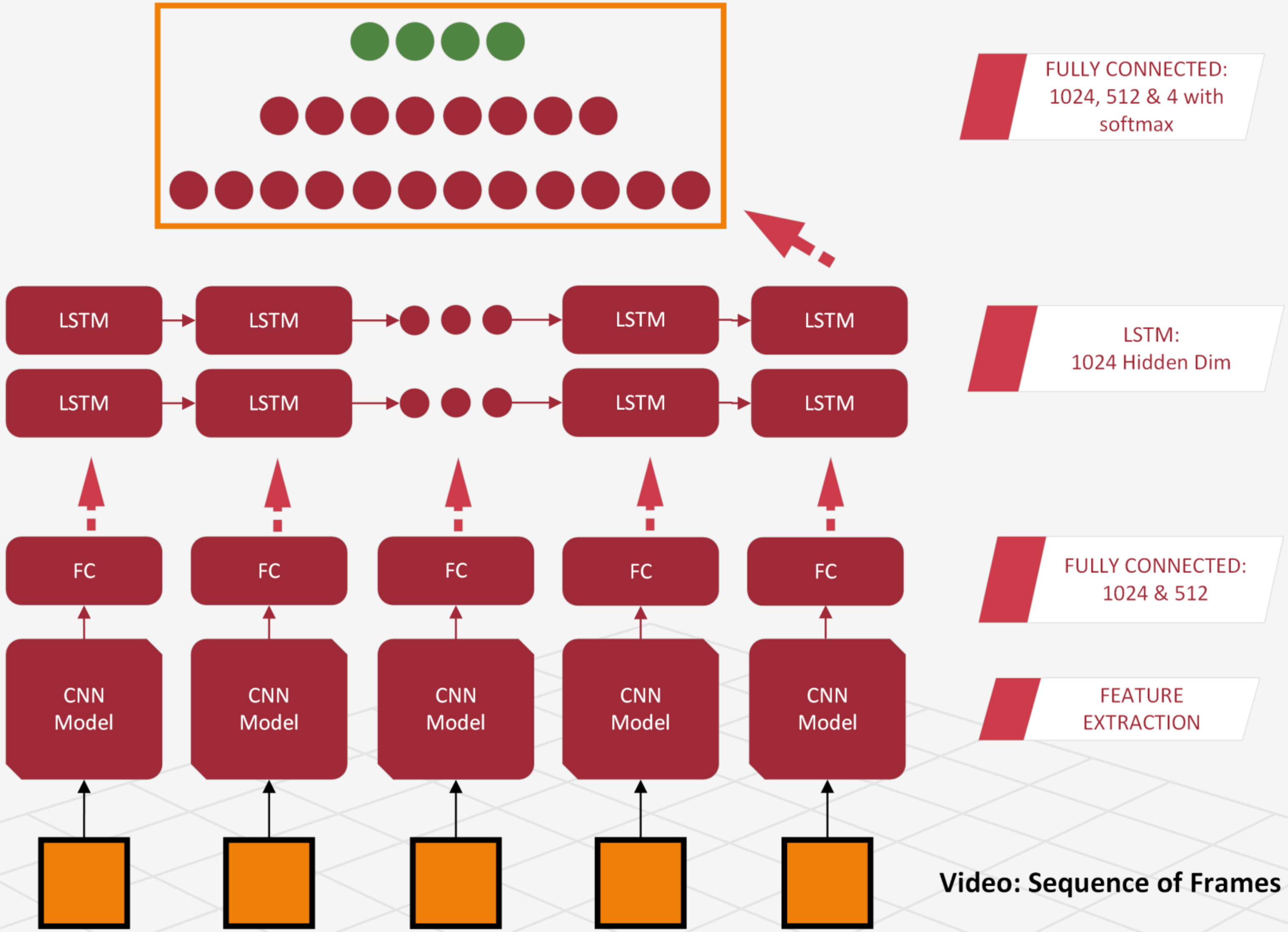

FULLY CONNECTED:
512 & 4 with softmax
LSTM
LSTM
LSTM
LSTM
LSTM:
1024 Hidden Dim
Feature Sequence
CNN Model
CNN Model
CNN Model
CNN Model
CNN Model
FEATURE EXTRACTION
Video: Sequence of Frames

Approach B



Video: Sequence of Frames

Approach C



FULLY CONNECTED: 1024, 512 & 4 with softmax

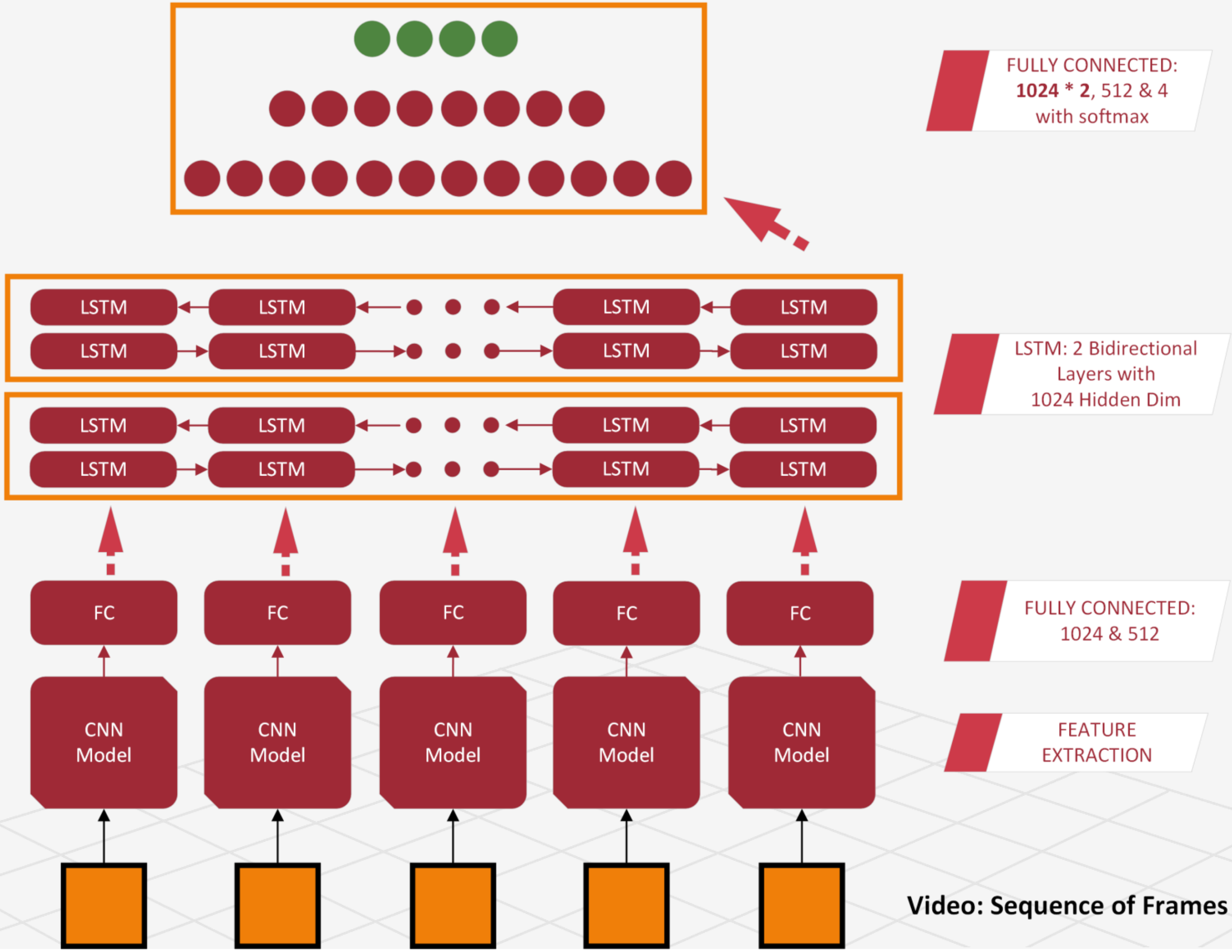
LSTM: 1024 Hidden Dim

FULLY CONNECTED: 1024 & 512

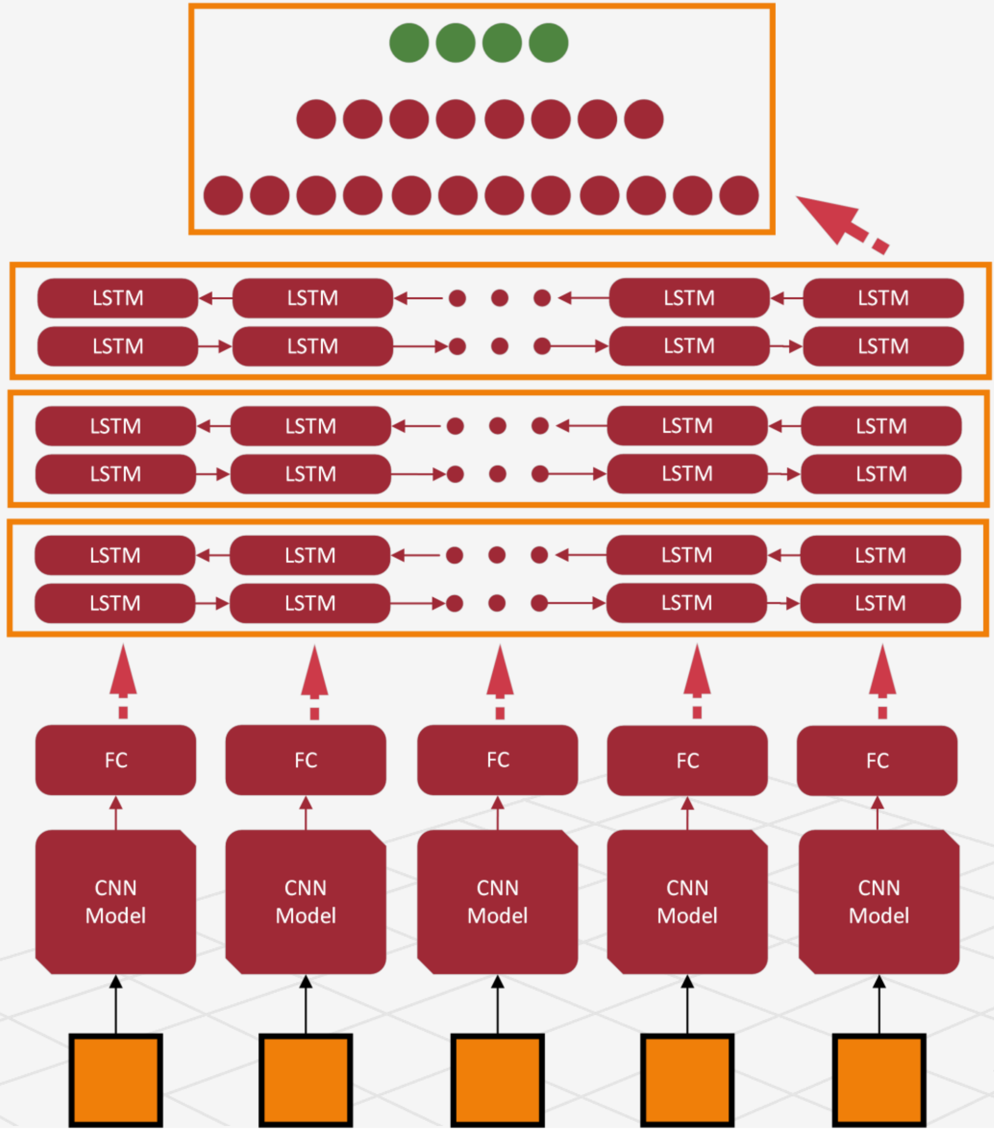
FEATURE EXTRACTION

Video: Sequence of Frames

Approach C’


FULLY CONNECTED:
1024 * 2, 512 & 4
with softmax
LSTM
LSTM: 2 Bidirectional
Layers with
1024 Hidden Dim
FC
FULLY CONNECTED:
1024 & 512
CNN
Model
FEATURE
EXTRACTION
Video: Sequence of Frames

Approach D



FULLY CONNECTED:
1024 * **2**, 512 & 4
with softmax

LSTM: 3 Bidirectional
Layers with
1024 Hidden Dim

FULLY CONNECTED:
1024 - 512 - 256

FEATURE
EXTRACTION

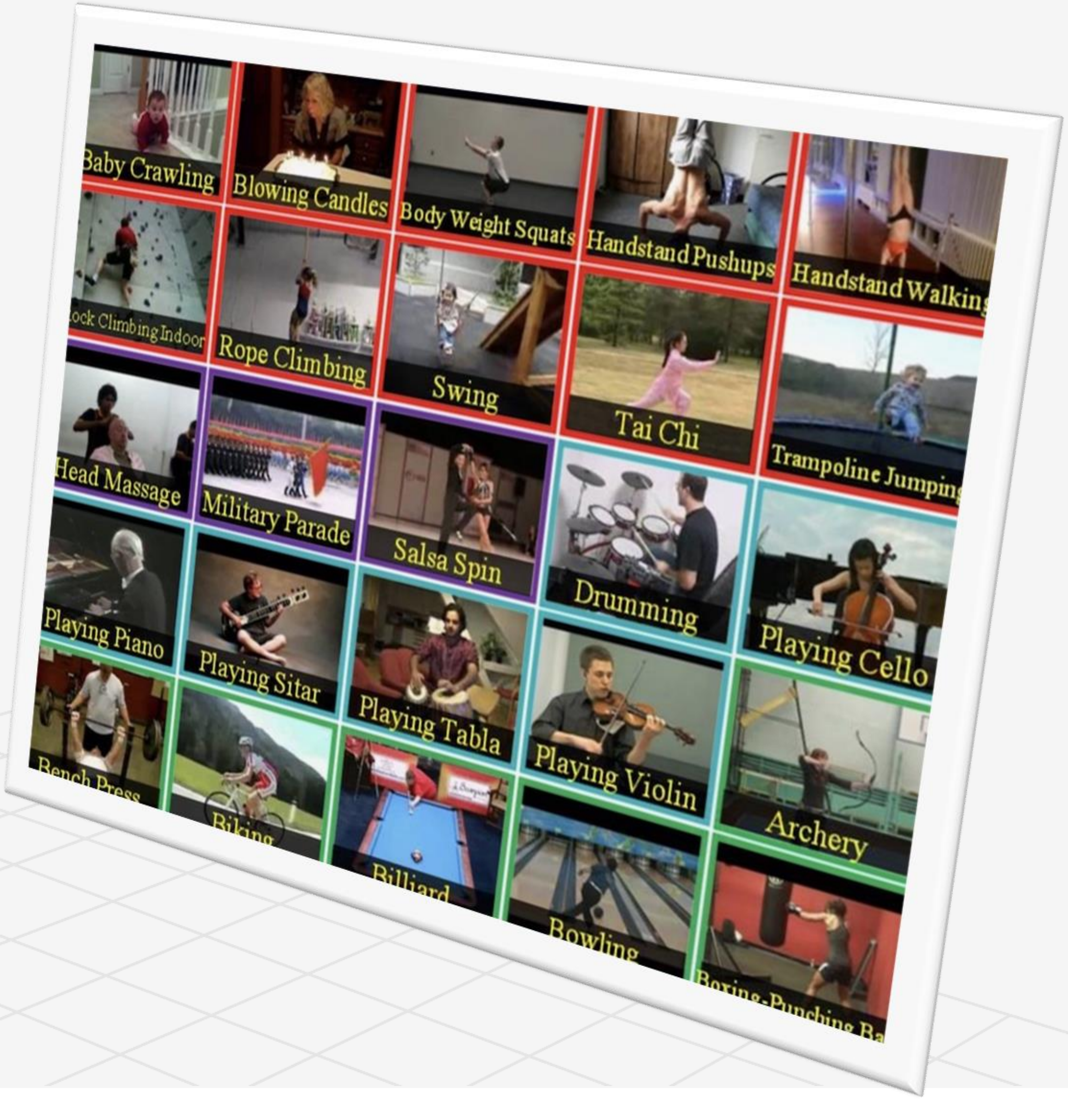
Video: Sequence of Frames

RESULTS

Dataset

UCF101 Human Actions dataset:

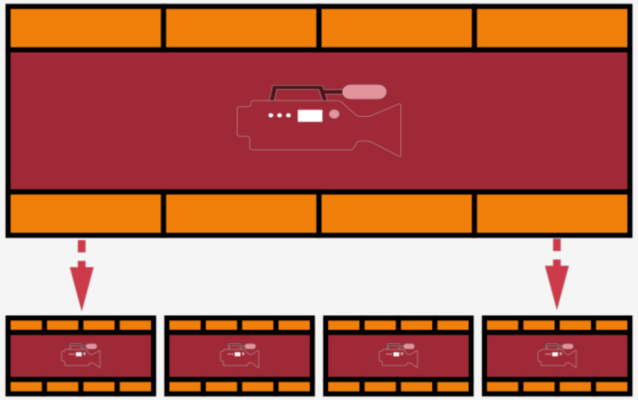
- A small subset was used in this project
- Official train-test splitting
- Classes:
  - 0: Playing Guitar
  - 1: Rock Climbing Indoor
  - 2: Soccer Juggling
  - 3: Band Marching



Preprocessing

Sequence level transformation:

- Problem: Differences in the total number of frames per video & the Fps
- Solution: Extract video segments → Shorter videos with fixed number of frames

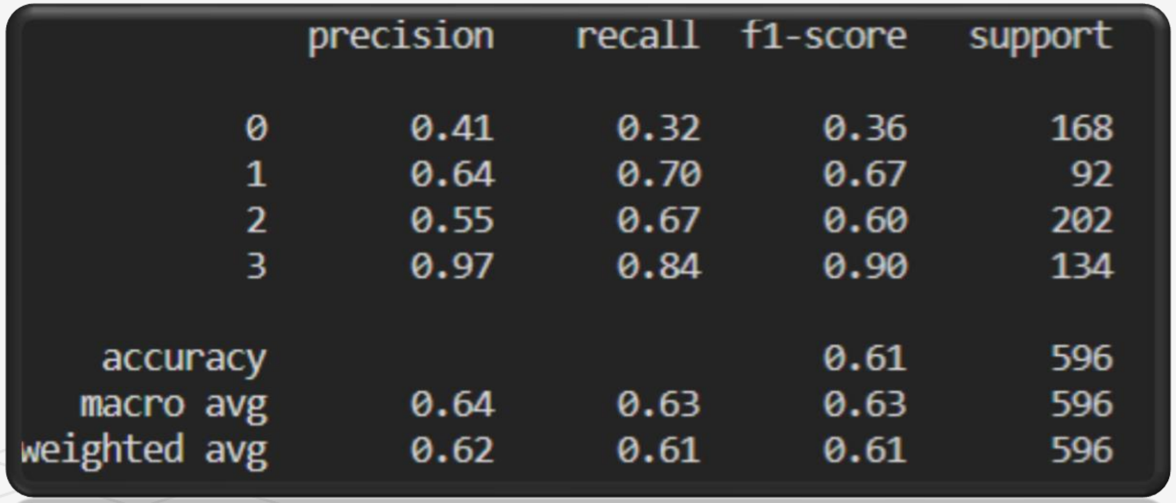


Frame-level transformation:

- Resize (e.g. 224x224)
- Normalize → *mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]*

Results: Baseline

True\Predicted	0	1	2	4
0	54	28	82	4
1	17	63	12	0
2	59	8	135	0
3	4	0	18	112



```
              precision    recall  f1-score   support

           0       0.41      0.32      0.36       168
           1       0.64      0.70      0.67        92
           2       0.55      0.67      0.60       202
           3       0.97      0.84      0.90       134

    accuracy                           0.61       596
   macro avg       0.64      0.63      0.63       596
weighted avg       0.62      0.61      0.61       596
```

Results

Pre-trained CNN selection:

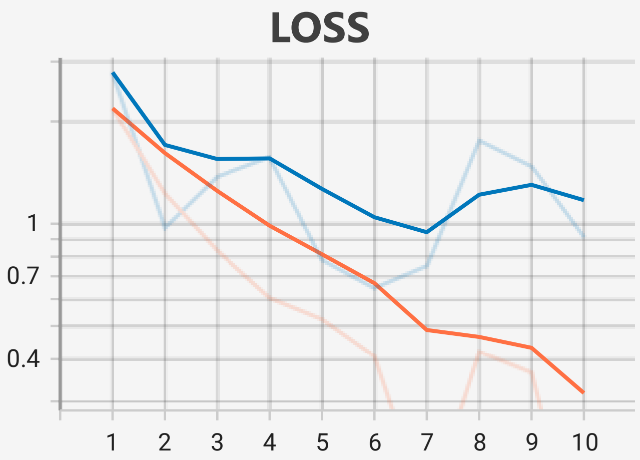
All models tested using the DNN of ***Approach A***

- ResNet – 125 freezed at the last Convolutional Layer (we receive output of Average Pooling) outperforms VGG 19 freezed at the penultimate Layer of classifier [97% vs 92% F1 Score in validation dataset].
- If we freeze the last Convolutional Layer of VGG the model can’t learn…
- MobileNet overfits…

Results

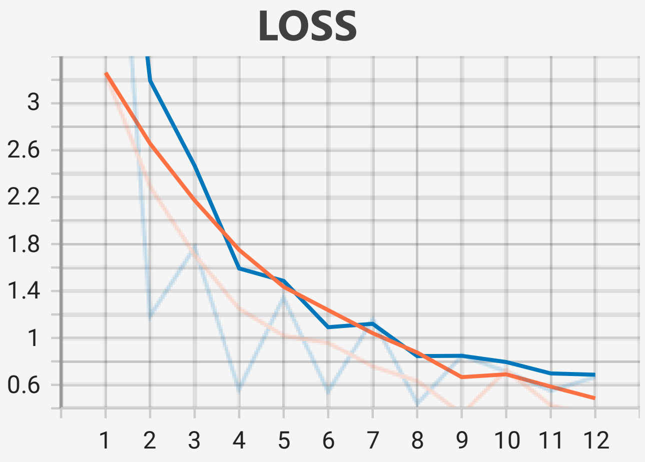
Model Name	F1 Score	Loss
Approach A	87.8	2.9
Approach B	97.7	0.64
Approach C	98.4	0.44
Approach C (Bid)	97.7	0.57
Approach D	98.6	0.29

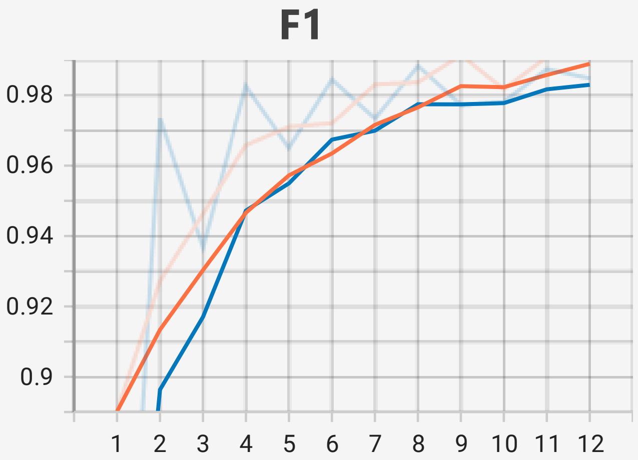
Results



Model Name	F1 Score	Loss
Approach A	87.8	2.9
Approach B	97.7	0.64
Approach C	98.4	0.44
Approach C (Bid)	97.7	0.57
Approach D	98.6	0.29

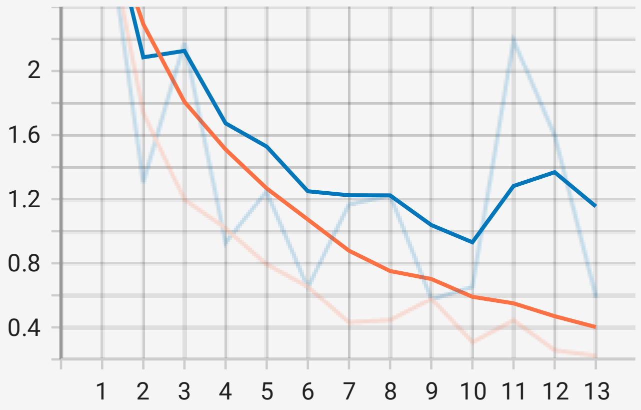
Results



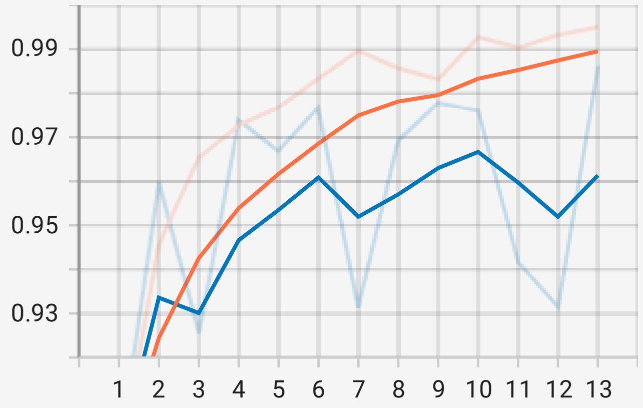


Model Name	F1 Score	Loss
Approach A	87.8	2.9
Approach B	97.7	0.64
Approach C	98.4	0.44
Approach C (Bid)	97.7	0.57
Approach D	98.6	0.29

Results



LOSS

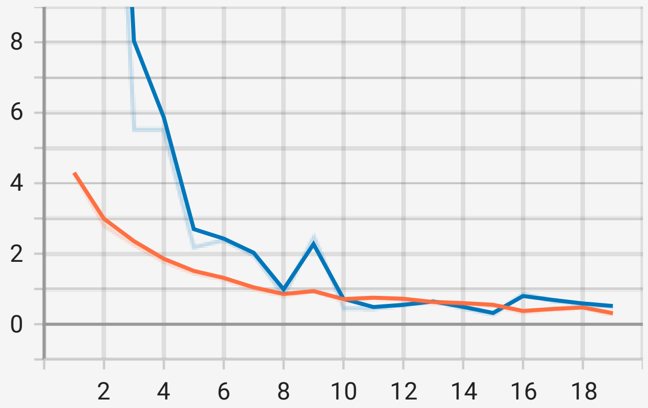


F1

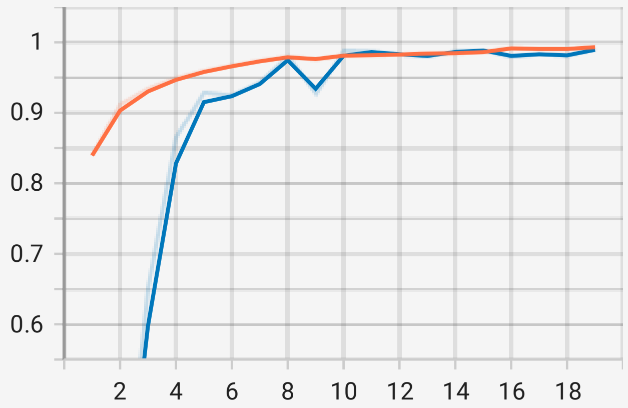
Model Name	F1 Score	Loss
Approach A	87.8	2.9
Approach B	97.7	0.64
Approach C	98.4	0.44
Approach C (Bid)	97.7	0.57
Approach D	98.6	0.29

Results

LOSS



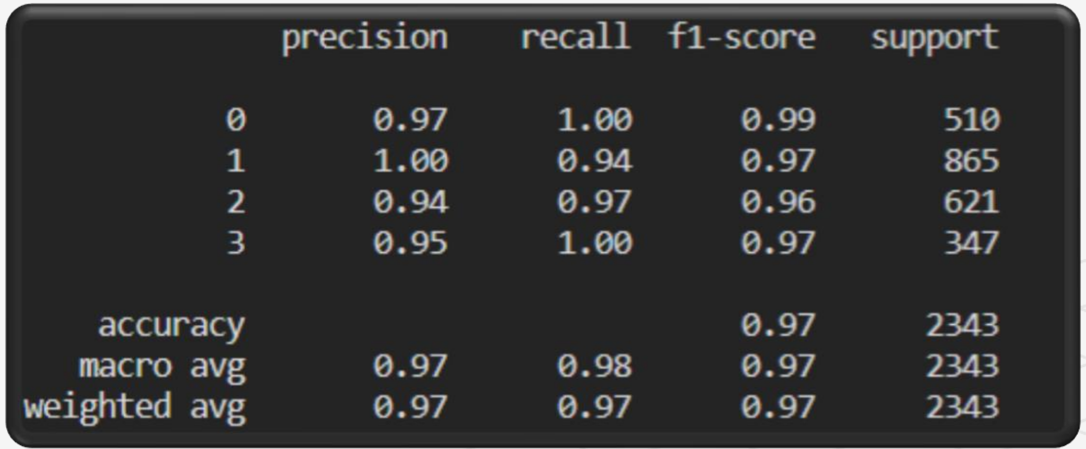
F1



Model Name	F1 Score	Loss
Approach A	87.8	2.9
Approach B	97.7	0.64
Approach C	98.4	0.44
Approach C (Bid)	97.7	0.57
Approach D	98.6	0.29

Results: Approach C & Test Dataset

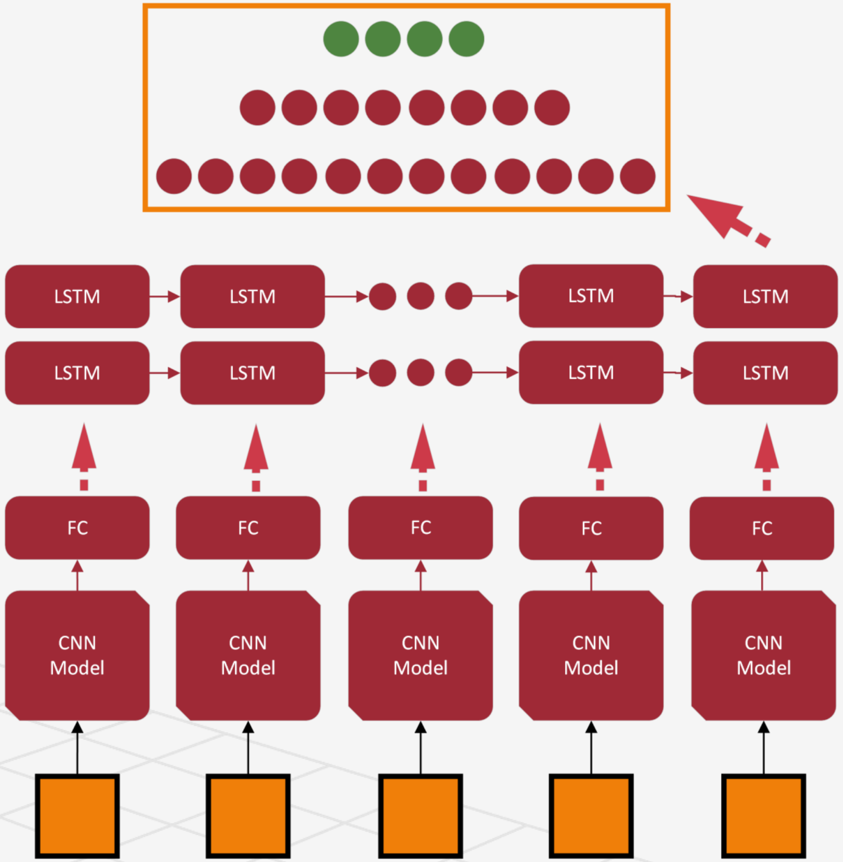
True\Predicted	0	1	2	4
0	510	0	0	0
1	2	812	37	14
2	13	1	603	4
3	0	1	0	346



```
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       510
           1       1.00      0.94      0.97       865
           2       0.94      0.97      0.96       621
           3       0.95      1.00      0.97       347

    accuracy                           0.97      2343
   macro avg       0.97      0.98      0.97      2343
weighted avg       0.97      0.97      0.97      2343
```



FULLY CONNECTED: 1024, 512 & 4 with softmax

LSTM: 1024 Hidden Dim

FULLY CONNECTED: 1024 & 512

FEATURE EXTRACTION

Video: Sequence of Frames

Thank you!