Highlights Detection in Sports Videos

Laboratory Supervisor:
Prof. Mathieu SALZMANN

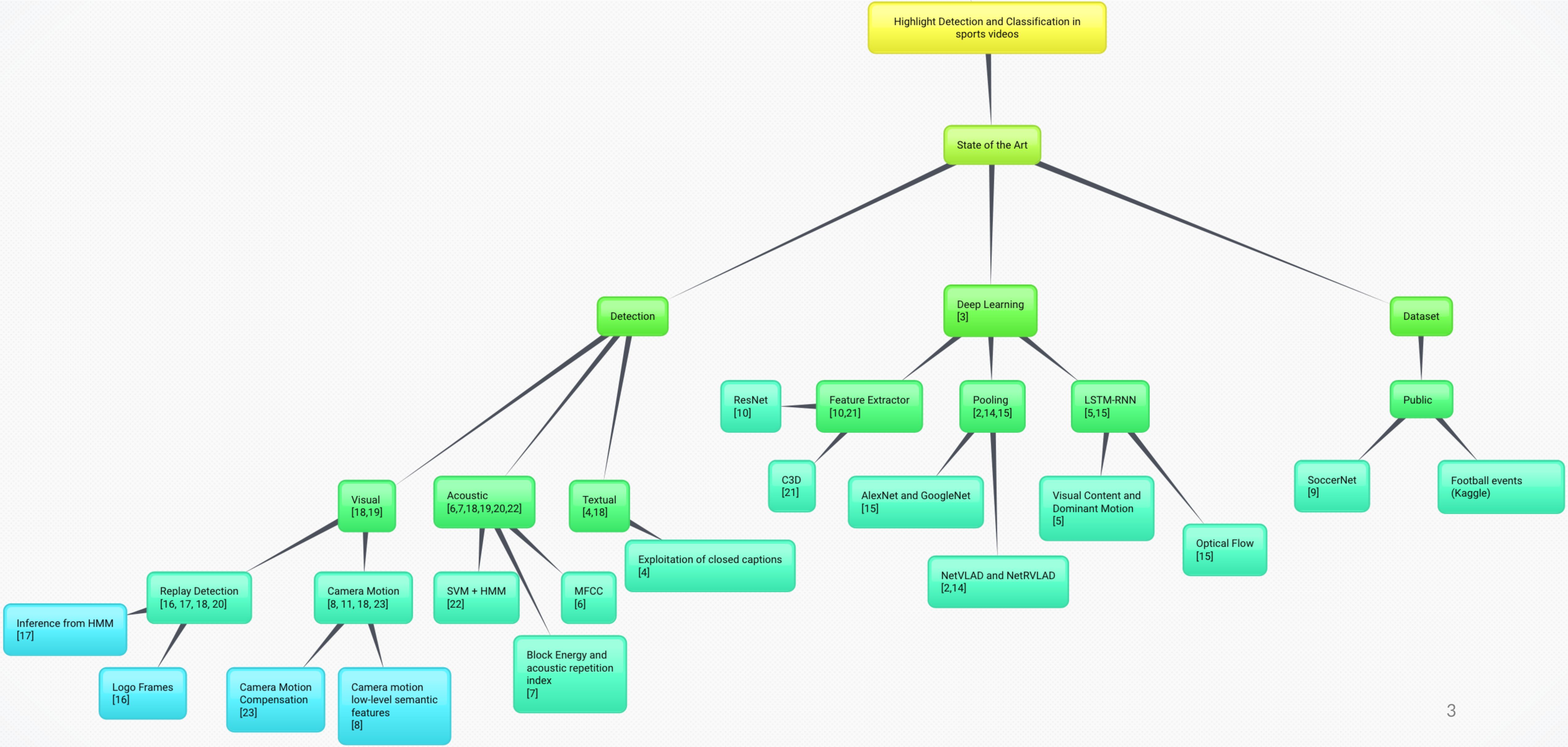
Industry Supervisor:
Mr. Alexandre ROUXEL

Author:
Vikalp KAMDAR

European Broadcasting Union

- Leading Expert in Media Broadcasting across Europe
- Alliance of 116 Public Media Services
- Over 2000 television, radio and online channels
- Partnership with T&I department and RTS

State of the Art



Highlight Detection and Classification in sports videos

State of the Art

Detection

Visual [18,19]

Replay Detection [16, 17, 18, 20]

Inference from HMM [17]

Logo Frames [16]

Camera Motion [8, 11, 18, 23]

Camera Motion Compensation [23]

Camera motion low-level semantic features [8]

Acoustic [6,7,18,19,20,22]

SVM + HMM [22]

Block Energy and acoustic repetition index [7]

MFCC [6]

Textual [4,18]

Exploitation of closed captions [4]

Deep Learning [3]

Feature Extractor [10,21]

ResNet [10]

C3D [21]

Pooling [2,14,15]

AlexNet and GoogleNet [15]

NetVLAD and NetRVLAD [2,14]

LSTM-RNN [5,15]

Visual Content and Dominant Motion [5]

Optical Flow [15]

Dataset

Public

SoccerNet [9]

Football events (Kaggle)

For this Project

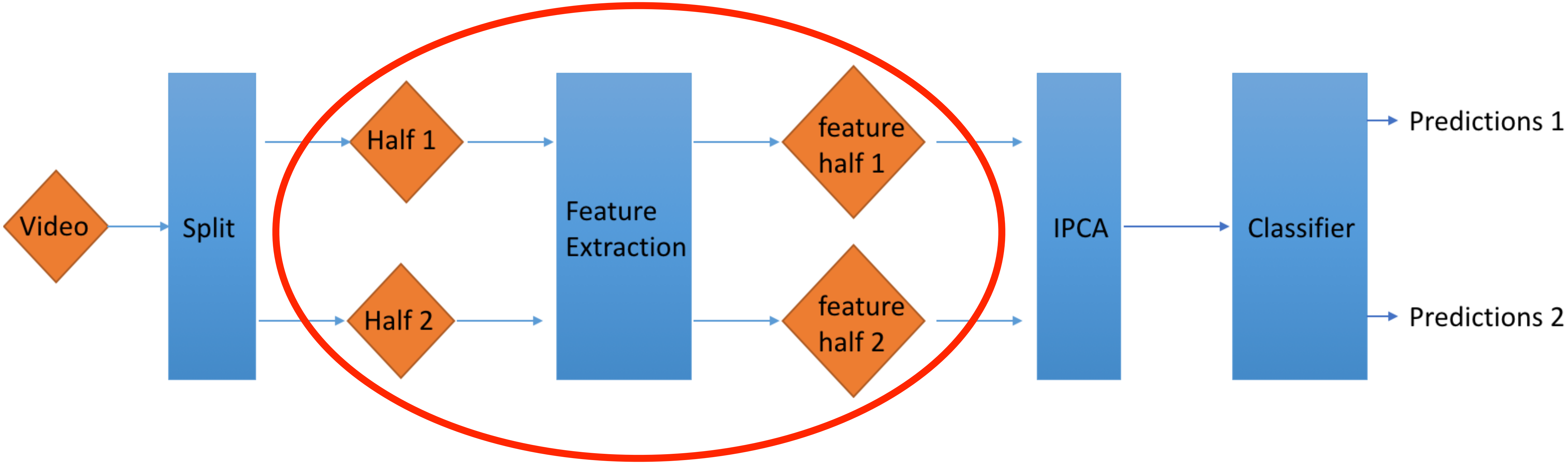
- Sport: Football

- Method: Deep Learning

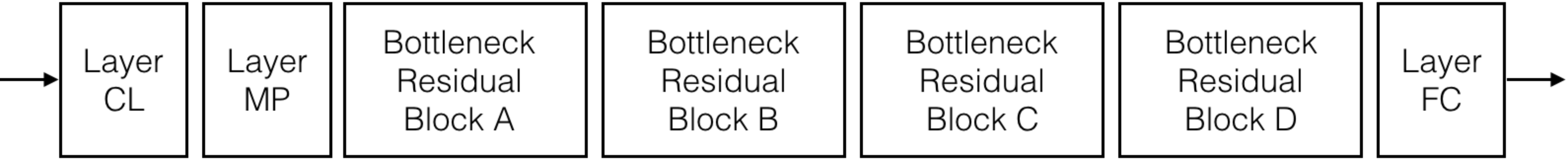
- Datasets:
    1. SoccerNet
    2. RTS

- Highlights:
    1. Cards
    2. Substitutions
    3. Goals

Pipeline


Video
Split
Half 1
Half 2
Feature Extraction
feature half 1
feature half 2
IPCA
Classifier
Predictions 1
Predictions 2

Feature extraction - ResNet (-152)


Layer CL
Layer MP
Bottleneck Residual Block A
Bottleneck Residual Block B
Bottleneck Residual Block C
Bottleneck Residual Block D
Layer FC

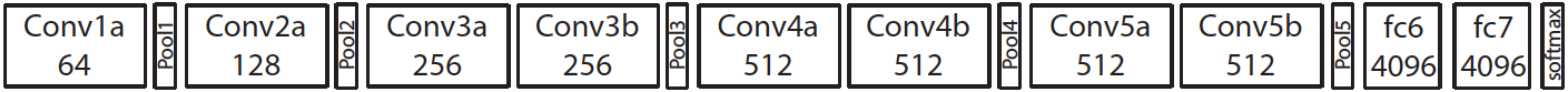
- Deep Residual Neural Network with 150 convolutional layers, split into 4 blocks, 1 max pooling layer and 1 fully connected layer

- Pre-Trained on ImageNet

- Outputs Spatial Features

- Feature Vector of dimension 2048

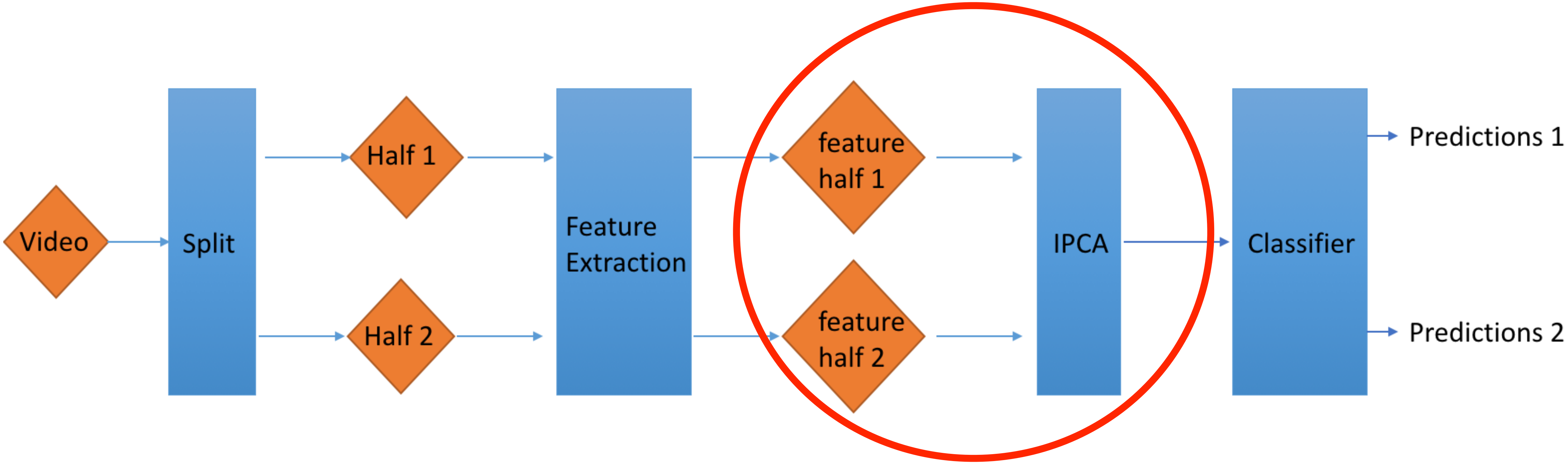
Feature extraction - C3D



Conv1a 64 | Pool1 | Conv2a 128 | Pool2 | Conv3a 256 | Conv3b 256 | Pool3 | Conv4a 512 | Conv4b 512 | Pool4 | Conv5a 512 | Conv5b 512 | Pool5 | fc6 4096 | fc7 4096 | softmax

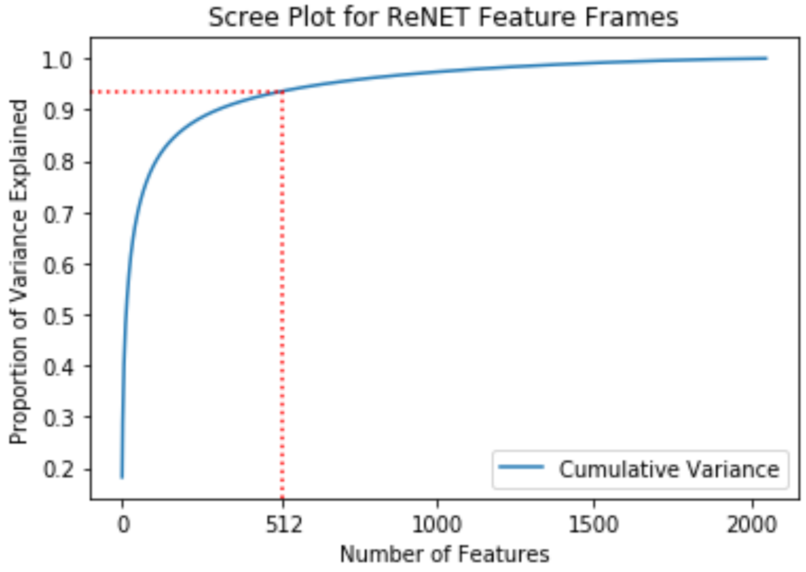
- Deep Convolutional Neural Network, implemented with 3D Convolutional and 3D Pooling Layers
- Pre-Trained on Sports1M
- Outputs spatio-temporal features
- Feature Vector of dimension 4096

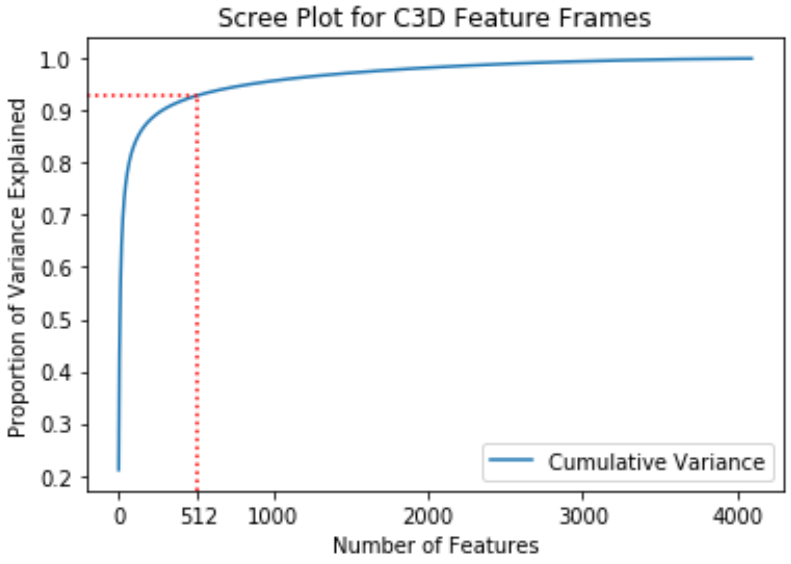
Pipeline


Video
Split
Half 1
Half 2
Feature Extraction
feature half 1
feature half 2
IPCA
Classifier
Predictions 1
Predictions 2

IPCA

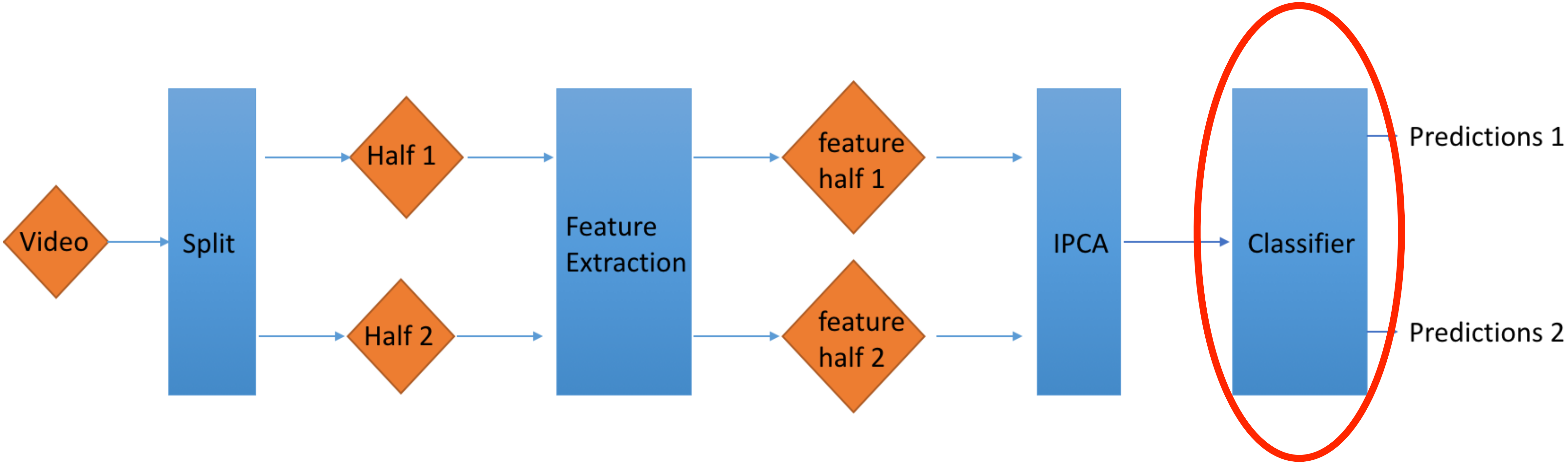
- Incremental Dimensionality Reduction: Incremental PCA
- Uniform Feature Vector dimension between ResNet and C3D



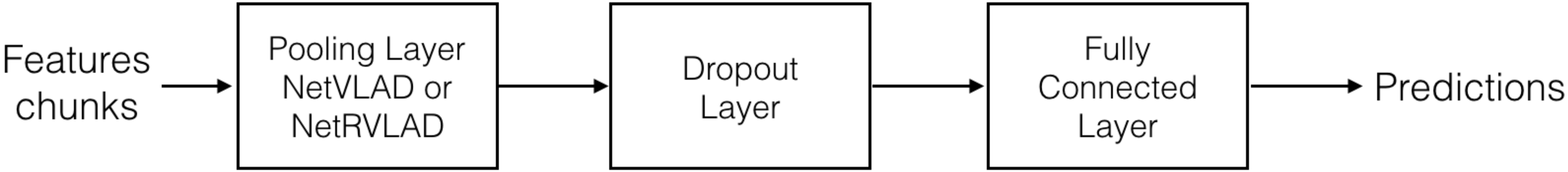


- 512 components explains
  - 93.65% of the total variance for ResNet on RTS dataset (94.29% on SoccerNet)
  - 92.94% of the total variance for C3D on RTS dataset (93.86% on SoccerNet)

Pipeline


Video
Split
Half 1
Half 2
Feature Extraction
feature half 1
feature half 2
IPCA
Classifier
Predictions 1
Predictions 2

Classifier



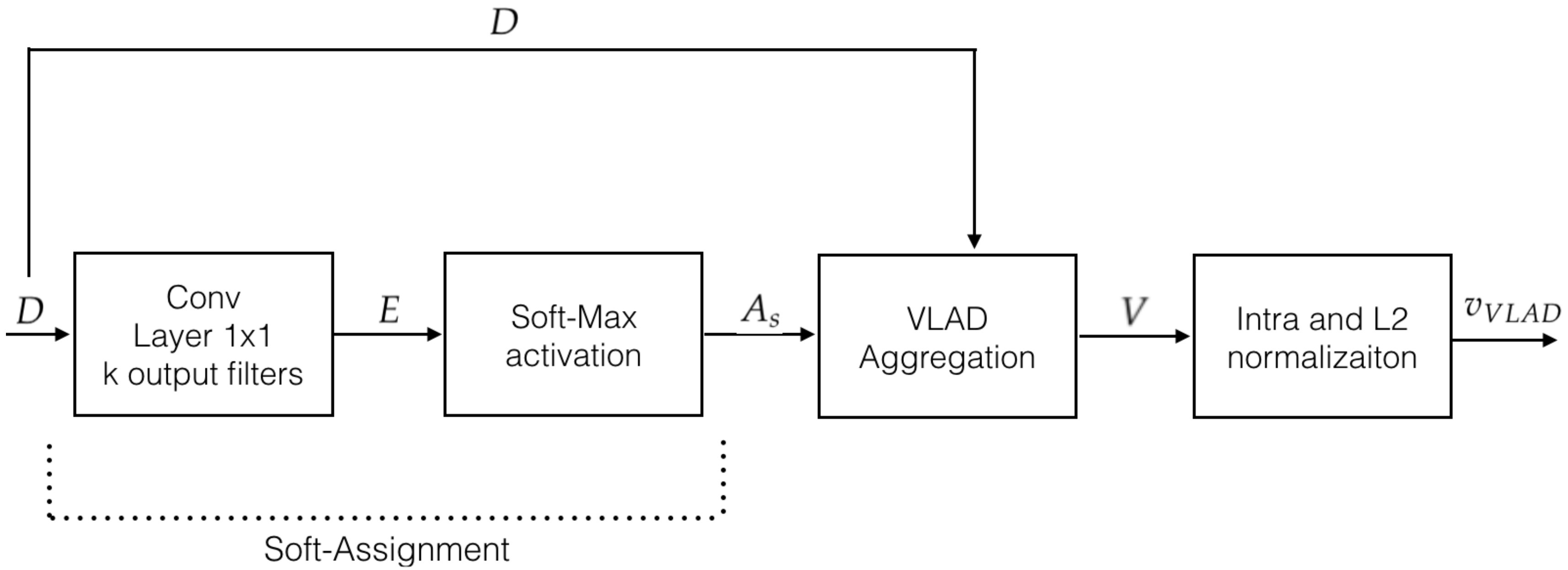
- Input: Extracted Features 1 minute chunks

- Output: Highlights Prediction Probabilities Peaks

- Adam Optimizer, to minimize a cross entropy loss function

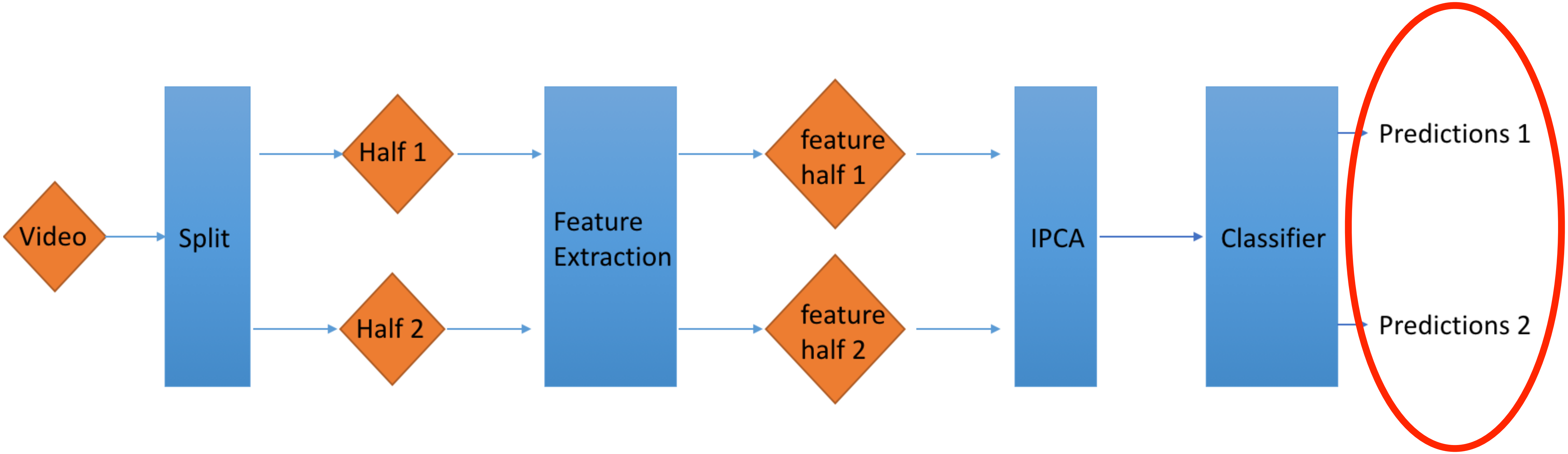
- Adaptive 0.01 learning rate

NetVLAD / NetRVLAD Pooling Layers



- End-to-end trainable layer based on VLAD / RVLAD descriptor pooling method
- Image Descriptors (Extracted Features) D, as input
- Convolutional Layer + Soft Max Activation = Soft Cluster Assignment
- Temporal Aggregation (Redundant information with C3D)
- k = 256 clusters (best performance SoccerNet)

Pipeline


Video
Split
Half 1
Half 2
Feature Extraction
feature half 1
feature half 2
IPCA
Classifier
Predictions 1
Predictions 2

Prediction Probabilities Peaks

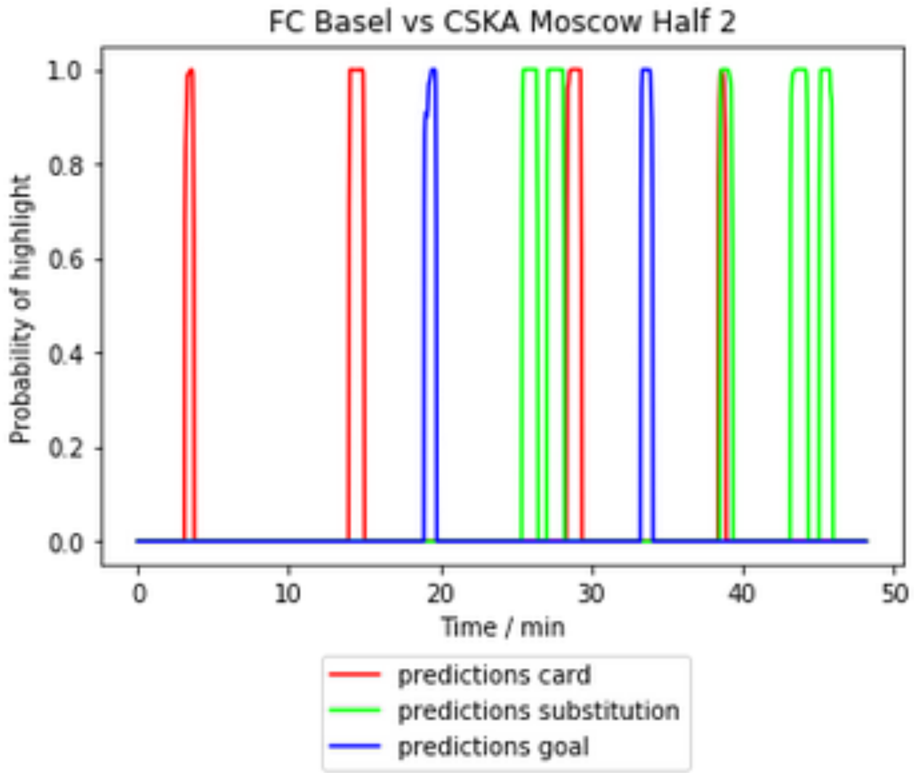
- An array P, where for each instant t, we have the probability that this instant corresponds to: No Highlight, Card, Substitution, Goal

$P = [p_1, p_2, ..., p_{T_s}]^T$, with $p_i = [Pr(No\ Highlight),\ Pr(Card),\ Pr(Substitution),\ Pr(Goal)]^T$.

- Smoothing Filter

- Thresholding at 0.6

- Duration of Highlight Clip:
  - Card and Substitution: 30s
  - Goal: 40s



Datasets

1. SoccerNet:
    - 500 Football matches
    - 6637 recorded highlights
    - Average of 1 highlight every 6.7 minutes

2. RTS:
    - 690 Football matches
    - 7254 recorded highlights
    - Hand Annotated highlights positions

Performances

- Overall ResNet with NetRVLAD outperforms C3D with NetVLAD
- Substitutions Highlights outperforms Goals and Cards
- Low Precision, Misclassifications of ‘Non-Highlights’ moments into Highlights