Course Project Demo
DA323: Multimodal Data Processing

Soccer Highlights Generator- Approach

Course Instructor: Dr. Neeraj Sharma

Name: Prakhar Punj
Roll No. : 220150011

Improved Soccer Action Spotting using both Audio and Video Streams

Introduction

Two tasks:

- Action classification
- Action spotting

Action spotting:

- Finding temporal anchors of events in a video
- Several issues:
  - No clear start and end frames for important actions
  - Actions are temporally discontinuous
  - Important actions are rare

SoccerNet dataset

SoccerNet is a benchmark dataset

- 500 soccer games from the Big Five European leagues
  - Training set: 300 games
  - Validation set: 100 games
  - Testing set: 100 games

- 6,637 events referenced split into 3 classes:
  - *goals*: the instant the ball crosses the goal line to enter the net
  - *cards*: the instant a card is shown by the referee
  - *subs*: the instant a new player enters the field to replace another one

- Addition of *background* class for the absence of the three events

Goals

Two tasks:

- Soccer action classification
- Soccer action spotting

Analyze benefit of audio stream:

- Should increase performance on both tasks
- Should provide useful information through sound events
  - Fans shout out when goal is scored
  - Red cards cause discontent

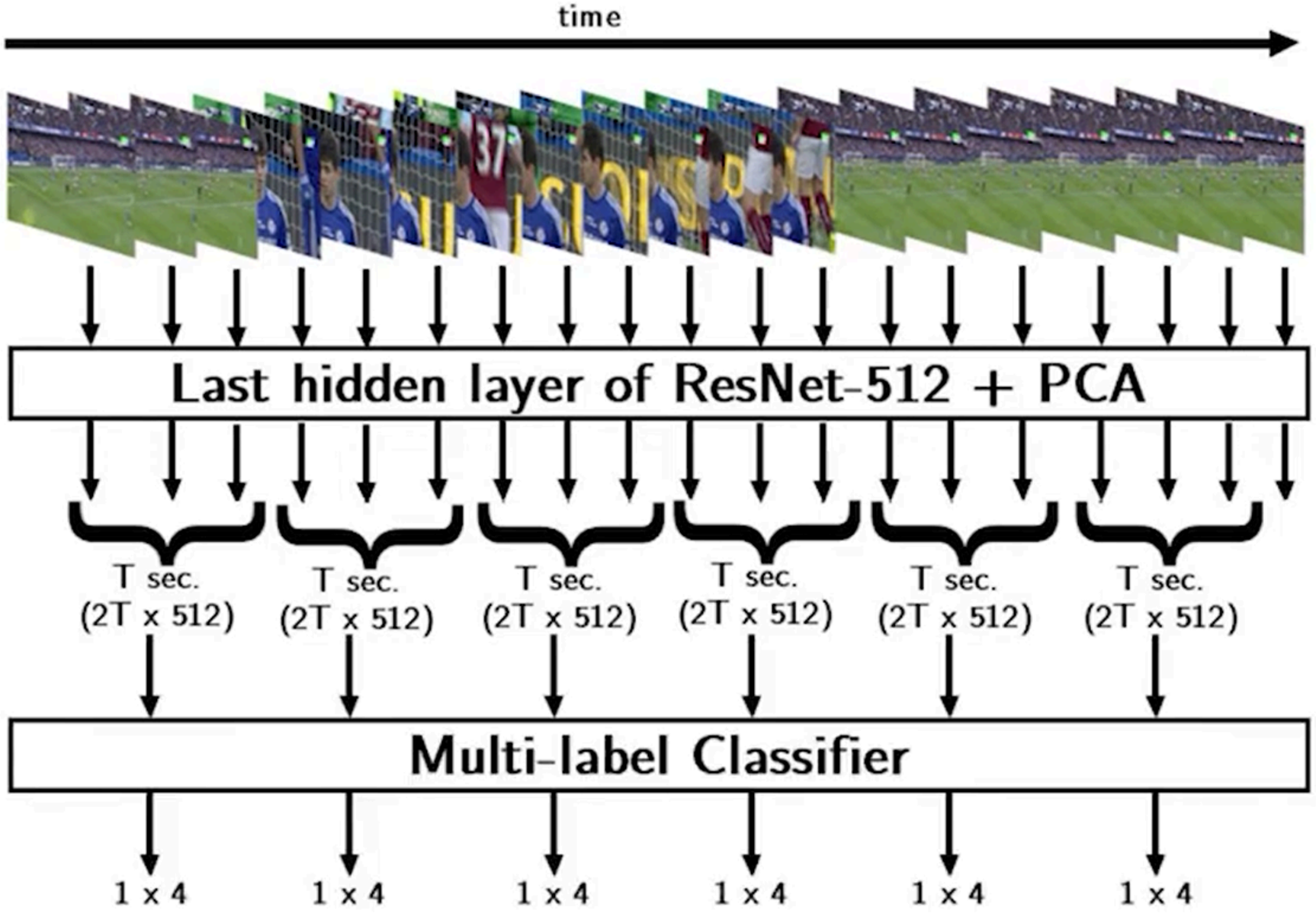
Video Representation

- Videos unified at 25 fps and trimmed at start of the game
- Frames sampled at 2 frames per second
- Sampled frames resized and cropped to 224×224 resolution
- ResNet-512 pretrained on 1000 categories ImageNet dataset
- Output from the last hidden layer
  $\longrightarrow$ 2048-dimensional vector for each frame
- Applied PCA to reduce dimension to 512 (retains 93.3% of variance)

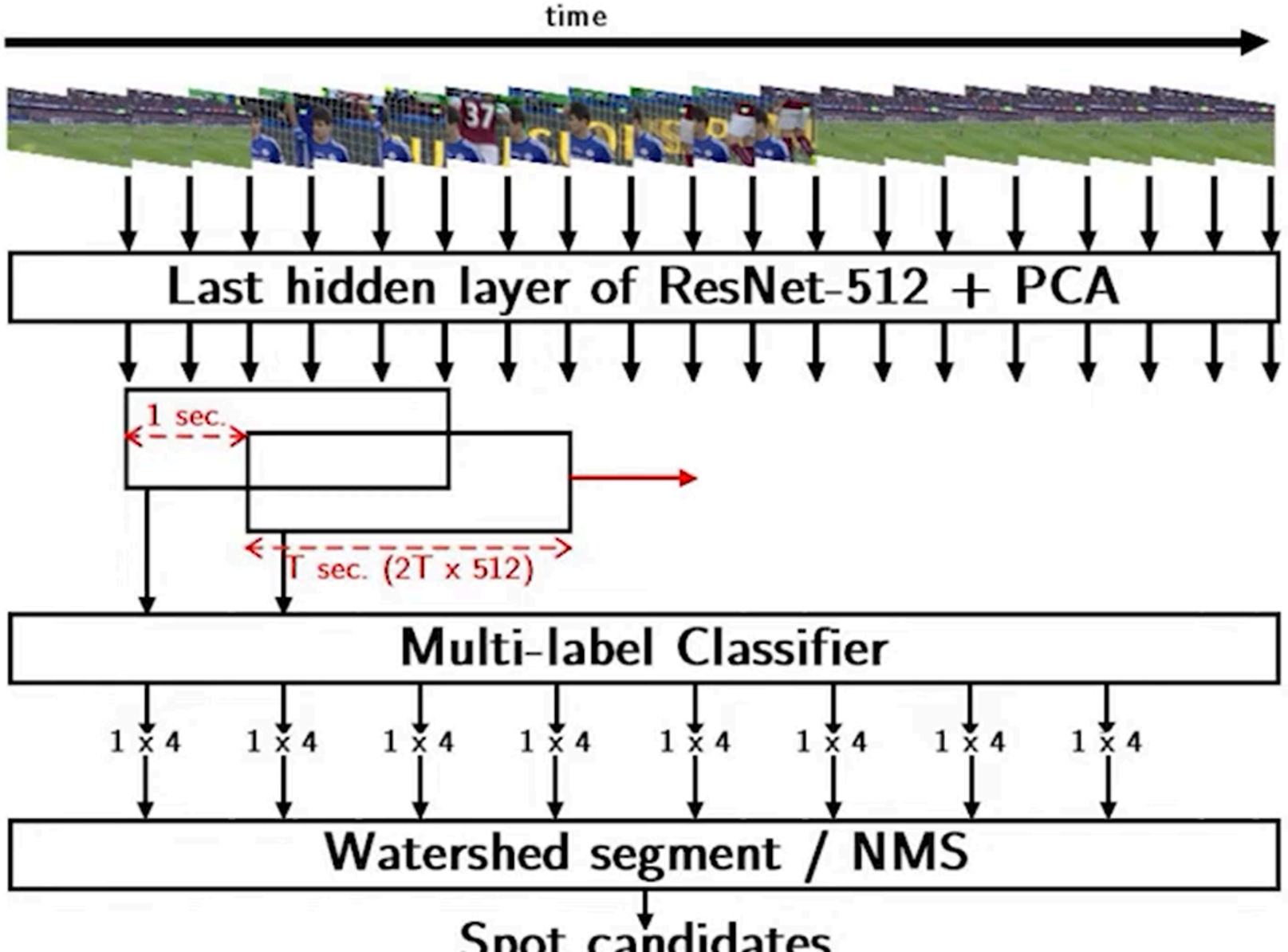
Audio Representation

- Audio streams trimmed at start of the game
- Audio streams divided into 0.5-second chunks
- VGGish pretrained on AudioSet dataset
- Output from the last convolutional layer
- Global average pooling
  $\longrightarrow$ 512-dimensional vector for each chunk

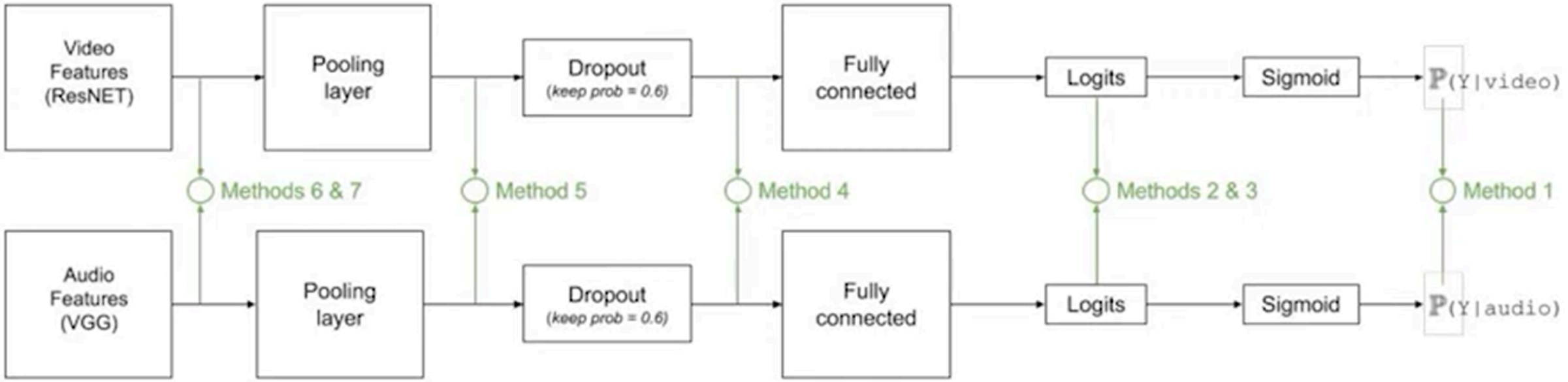
General approach - Action classification



General approach - Action spotting



Multimodal fusion



Temporal Pooling layer :

- NetRVLAD
- NetVLAD

Video chunk classification

- Metric used: mean Average Precision (mAP)

- Models used: best models of SoccerNet baseline
  - NetVLAD with $k = 512$ clusters
  - $k = 512$ high computation load
    $\Longrightarrow$ NetRVLAD with $k = 64$ to find best merging method

Video chunk classification - Merging methods

- NetRVLAD with $k = 64$ clusters

Models	$T = 60$ sec.	$T = 30$ sec.
Video baseline	66.0	58.7
Audio baseline	50.6	43.7
Merging method 1	68.4	63.7
Merging method 2	72.6	67.3
Merging method 3	73.4	69.3
Merging method 4	<u>73.7</u>	68.8
Merging method 5	72.8	68.7
Merging method 6	64.1	59.6
Merging method 7	64.2	58.1

Best merging method

- NetRVLAD with $k = 64$ clusters
- $T = 60$ sec.

Labels	**Video baseline**	**Audio baseline**	**Merge method 4**
"background"	97.6	96.7	98.0
"cards"	60.5	19.2	63.9
"substitutions"	69.8	55.1	72.6
"goals"	67.7	77.3	84.5

- Multimodal: better on all classes
- Audio only:
  - worst on all classes, except *goals*
  - carries information about *substitutions*
  - poor results for *cards*

Video chunk classification with best model

- Best model of SoccerNet baseline: NetVLAD, with $k = 512$ clusters
- Comparison with two window sizes: $T = 60$ sec. and $T = 20$ sec.

Models	$T = 60$ **sec.**	$T = 20$ **sec.**
Video-based NetVLAD baseline	67.5	56.6
Audio-based NetVLAD baseline	46.8	35.9
Audio + Video NetVLAD	75.2	75.0

$\longrightarrow$ same observations as with NetRVLAD, with $k = 64$ clusters

$\Longrightarrow$ In average, mAP increased by 7.43% by adding audio as input

Action spotting

- δ: tolerance in the precise time instant of the detected event
- Metric used: Average-mAP
  $\longrightarrow$ area under the mAP curve as a function of δ ranging from 5 to 60 seconds
- Evaluated models:
  - NetRVLAD, $k = 64$ clusters, $T = 60$ sec.
  - NetVLAD, $k = 512$ clusters, $T = 60$ sec.
  - NetVLAD, $k = 512$ clusters, $T = 20$ sec.

Action spotting - Results

Table 5. Average-mAP for action spotting.

	Video-only			Audio-only			Audio + Video		
Models	**Seg. max**	**Seg. center**	**NMS**	**Seg. max**	**Seg. center**	**NMS**	**Seg. Max**	**Seg. center**	**NMS**
NetRVLAD	30.8%	41.9%	30.2%	21.8%	30.3%	22.1%	34.0%	47.6%	33.4%
60-sec. chunks NetVLAD	29.6%	43.4%	29.0%	19.9%	27.1%	19.5%	32.3%	48.7%	31.8%
20-sec. chunks NetVLAD	49.2%	<u>50.2%</u>	49.4%	30.0%	<u>31.0%</u>	30.0%	54.0%	<u>56.0%</u>	53.6%

$\Longrightarrow$ In average, Average-mAP increased by 4.19% by adding audio as input

Conclusion

- The influence of audio stream was studied on two tasks:
  - Action classification
  - Action spotting
- Only audio stream is worse, except for the *goals* class
- Combining video and audio streams improves the performance on SoccerNet dataset:
  - Better results for every class
  - Increase performance of action classification by 7.43%
  - Increase performance of action spotting by 4.19%
- Smaller video chunk sizes perform:
  - worse for action classification
  - better for action spotting

Thank You

Course Instructor: Dr. Neeraj Sharma

Name: Prakhar Punj
Roll No. : 220150011