First Person Action Recognition: Temporal and Motion Approach

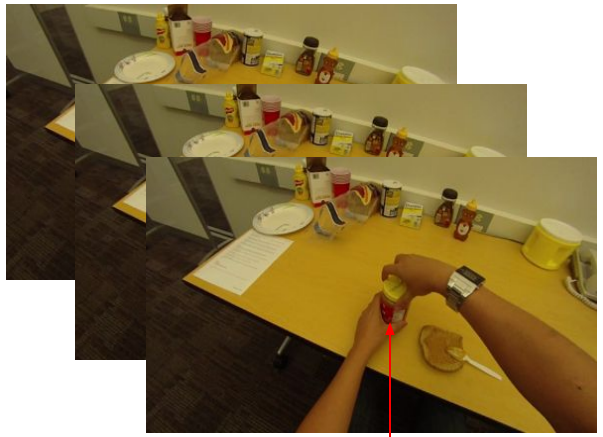
Valeria Sorrenti
Alessio Mongoli
Simone Santia

POLITECNICO
DI TORINO

First Person Action Recognition

- What is action recognition?
  - Identification of different actions from videoclips

- Differences between third and first person with problems
  - Camera Motion
  - Strong Occlusion
  - The amount of data



What is?

Dataset

GTEA-61:

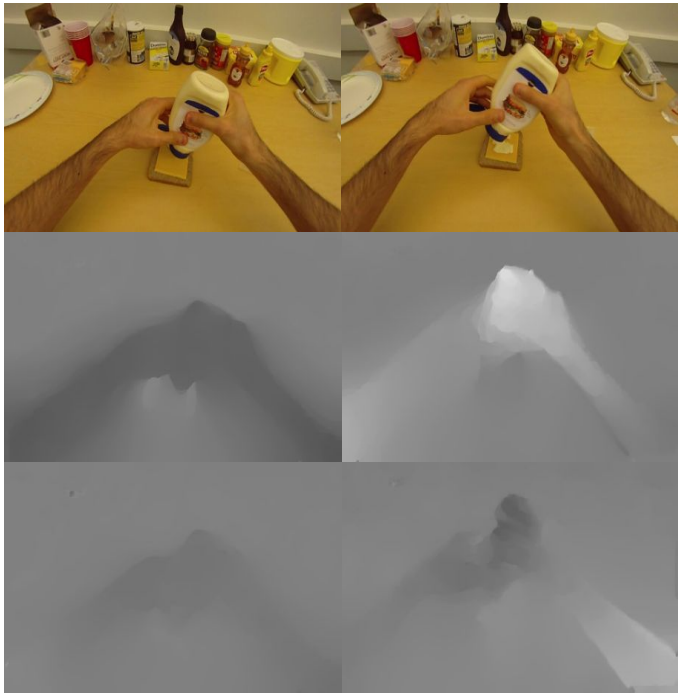
- 4 users: S1, S2, S3, S4
- 61 actions for each user
- Training sets = S1 S3 S4
- Validation set = test set = S2.

Optical flow?

a

b

c



Gtea-61 dataset: **a)**Rgb images; **b**)-**c)** Warp flow images(the camera motion subtracted from the optical flow)

Optical Flow

Optical flow is an image that tell us what is the pixel that move insight two different frames changing the brightness.

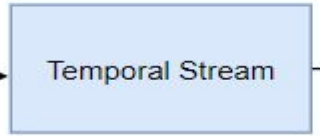
Networks

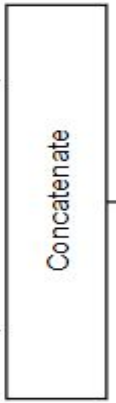
- Ego RNN
- SparNet
- Order Prediction task

Ego RNN - Two Stream

Rgb frames

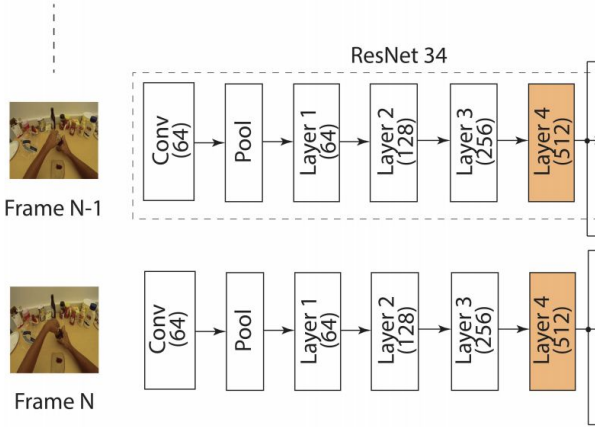
Stacked optical flow images





[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

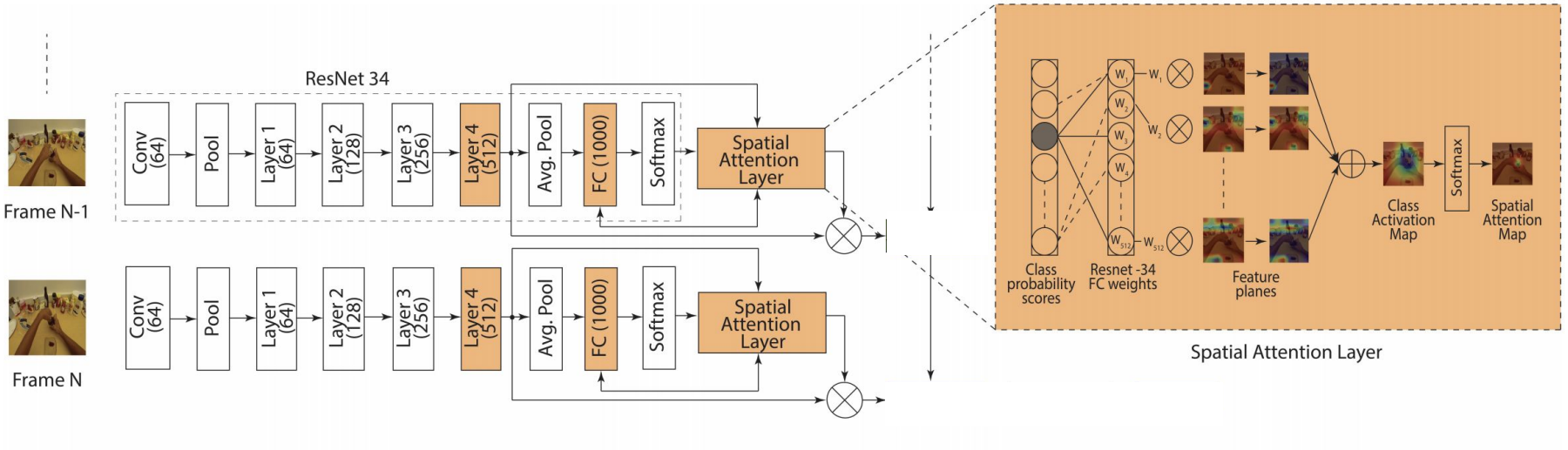
Ego RNN[*] - RGB



- Action recognition block (CNN backbone)
  - ResNet 34
  - 21M parameters
  - pretrained for generic image recognition
  - From each frame extract appearance features

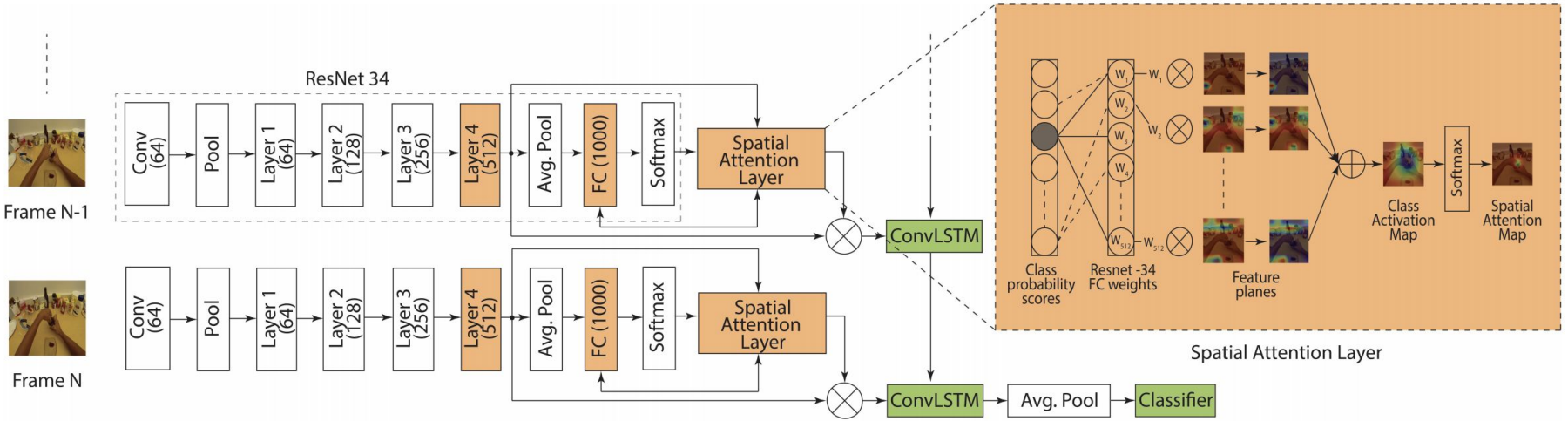
[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Ego RNN - RGB



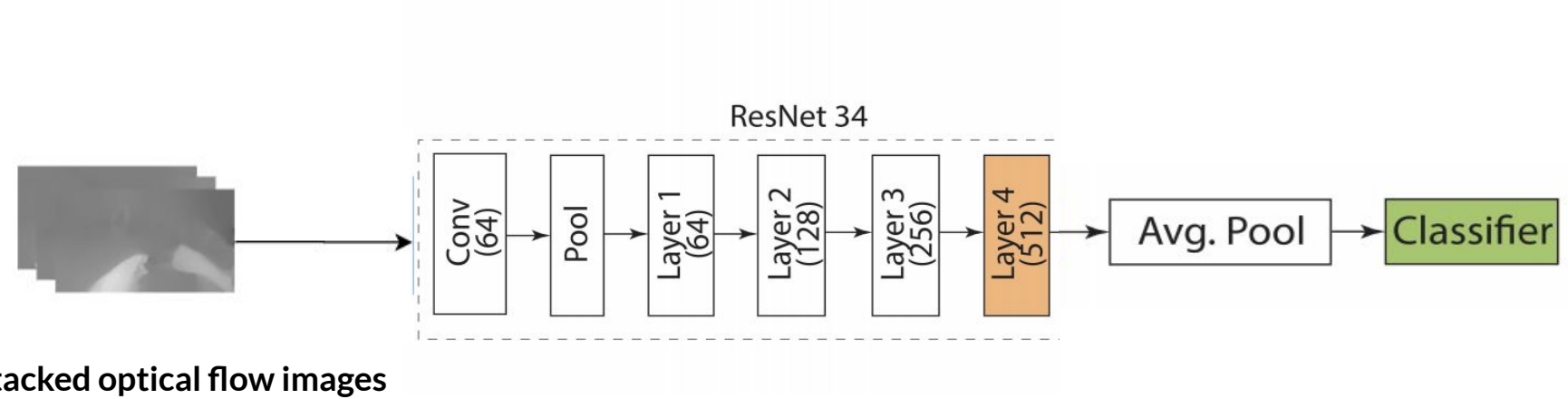
[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Ego RNN - RGB



[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Ego RNN - Temporal network



Stacked optical flow images

Ego RNN - Training Step

- The Ego RNN network is trained in 4 stage because the two network are trained separately

[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Ego RNN - Training Step

- The Ego RNN network is trained in 4 stage because the two network are trained separately
  - The spatial network is trained in two stage
    - First stage: we train from scratch only the classifier and the ConvLSTM
    - Second stage: we train the last convolutional layer and the Fully connected layer of the Resnet in addition to the convLSTM and classifier layer.

[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Ego RNN - Training Step

- The Ego RNN network is trained in 4 stage because the two network are trained separately
  - The spatial network is trained in two stage
    - First stage: we train from scratch only the classifier and the ConvLSTM
    - Second stage: we train the last convolutional layer and the Fully connected layer of the Resnet in addition to the convLSTM and classifier layer.
  - The Temporal network is trained in one stage using 5 stacked warp flow images

[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Ego RNN - Training Step

- The Ego RNN network is trained in 4 stage because the two network are trained separately
  - The spatial network is trained in two stage
    - First stage: we train from scratch only the classifier and the ConvLSTM
    - Second stage: we train the last convolutional layer and the Fully connected layer of the Resnet in addition to the convLSTM and classifier layer.
  - The Temporal network is trained in one stage using 5 stacked warp flow images
  - We concatenate the output of the two network by adding a FC on top of the two networks

[*] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Result

Configurations	Accuracy% 7 Frames	Accuracy% 16 Frames
ConvLSTM	45,55	52,83
ConvLSTM-attention	47,41	54,72
Two-stream (joint train)	56,03	**70,70**

	Accuracy% 5 Frames
Temporal-warp flow	44,83

Consideration

From the table of result we can see that:

- The spatial attention improves the performance
- Better results are obtained with 16 frames
- The highest result is achieved with the two stream Network

Ego RNN: issues

Main issues of two-stream networks:

- appearance and motion features are learned separately, sub-optimal solution
- growth of overall architecture and parameters to optimize, optimization is performed in multiple stages

Ego RNN: issues

Main issues of two-stream networks:

- appearance and motion features are learned separately, sub-optimal solution
- growth of overall architecture and parameters to optimize, optimization is performed in multiple stages

IDEA!

Ego RNN: issues

Main issues of two-stream networks:

- appearance and motion features are learned separately, sub-optimal solution
- growth of overall architecture and parameters to optimize, optimization is performed in multiple stages

IDEA!

SINGLE-STREAM NETWORKS WITH SELF-SUPERVISED TASKS

SparNet[*]: single-stream and self-supervised task

- Single-stream network: hold appearance stream

[*]Mirco Planamente, Andrea Bottino, Barbara Capputo, “Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition”

SparNet[*]: single-stream and self-supervised task

- Single-stream network: hold appearance stream
- insertion of self-supervised task
    - **M**otion **S**egmentation block: replace optical flow stream

[*]Mirco Planamente, Andrea Bottino, Barbara Capputo, “Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition”

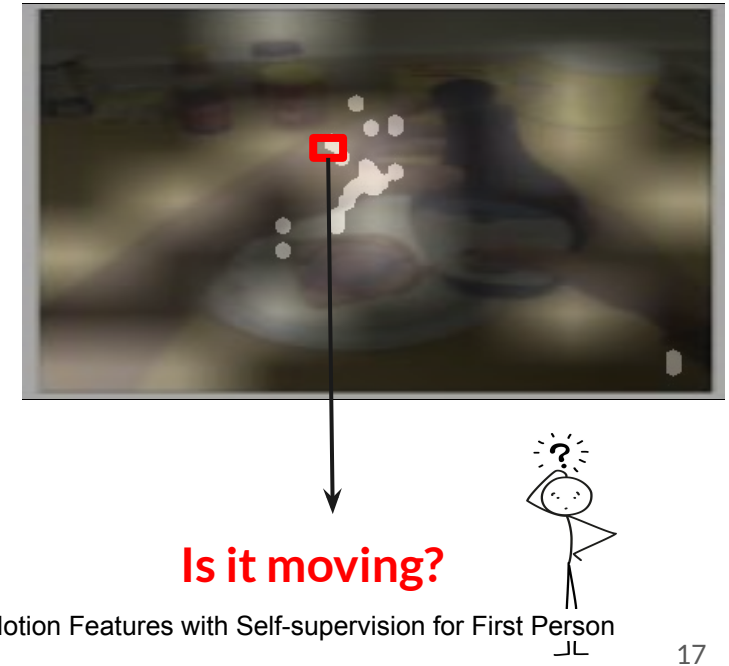
SparNet[*]: single-stream and self-supervised task

- Single-stream network: hold appearance stream
- insertion of self-supervised task
    - **M**otion **S**egmentation block: replace optical flow stream

[*]Mirco Planamente, Andrea Bottino, Barbara Capputo, “Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition”

SparNet[*]: single-stream and self-supervised task

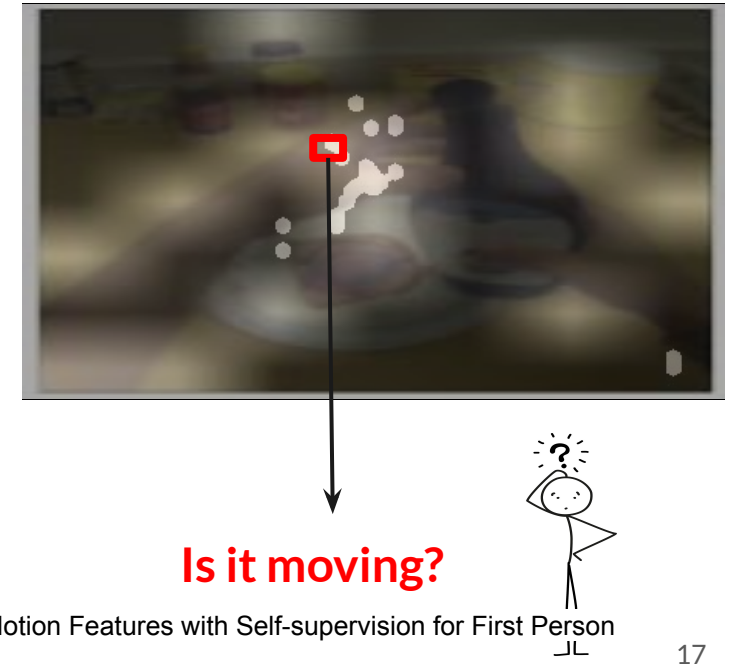
- Single-stream network: hold appearance stream
- insertion of self-supervised task
  - **M**otion **S**egmentation block: replace optical flow stream



[*]Mirco Planamente, Andrea Bottino, Barbara Capputo, “Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition”

SparNet[*]: single-stream and self-supervised task

- Single-stream network: hold appearance stream
- insertion of self-supervised task
  - **M**otion **S**egmentation block: replace optical flow stream

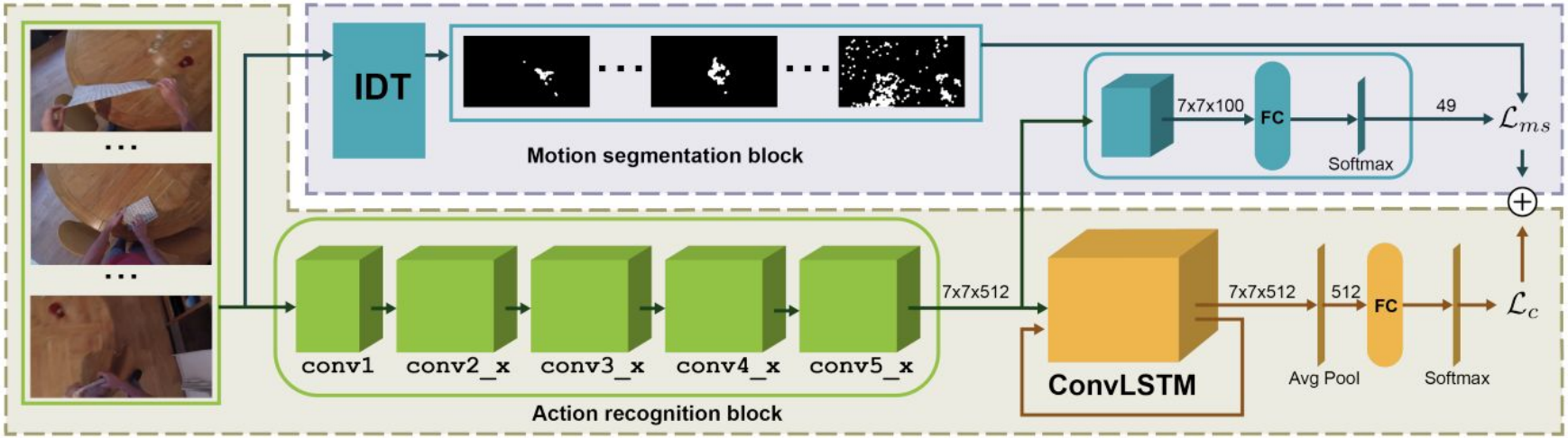


Two kind of implementation:

- Classification
- Regression

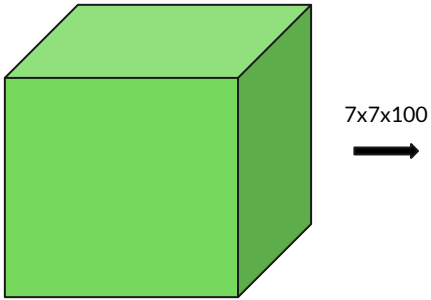
[*]Mirco Planamente, Andrea Bottino, Barbara Capputo, “Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition”

SparNet[*]: architecture overview

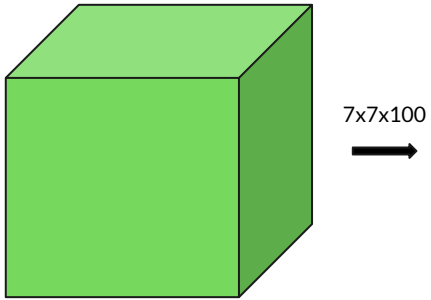


[*]Mirco Planamente, Andrea Bottino, Barbara Capputo, “Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition”

MS: Classification vs Regression

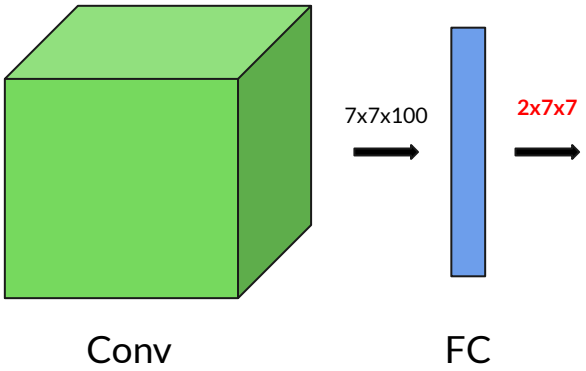


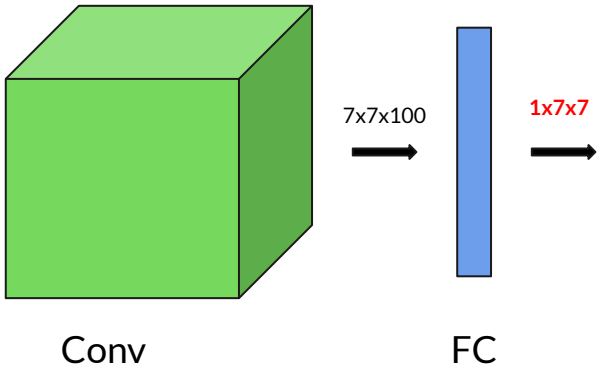
Conv



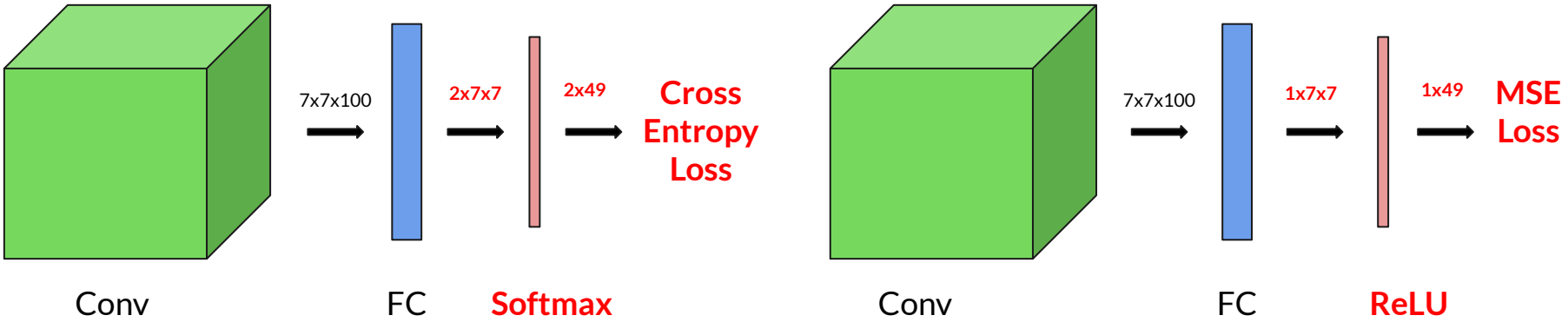
Conv

MS: Classification vs Regression





MS: Classification vs Regression



Results

	Accuracy % 7 frames		Accuracy % 16 frames	
	Regression	**Classification**	**Regression**	**Classification**
ConvLSTM	50,86	49,27	56,60	56,66
ConvLSTM-Attention	49,13	46,96	55,54	**59,43**

Parameters:

- **epochs:** 150
- **LR**: 1e-4
- **batch size**: 64
- **weight decay**: 4e-5

Fine-tuning:

- pretrained on ImageNet
- **first stage**: only classifier and ConvLSTM are trained

Augumentation:

- MultiscaleCornercrop
- RandomHorizontalFlip

SparNet: hyperparameters optimization

In order to find best results we trained model with 3 set of hyperparameters:

- **LR**: 1e-5, **Frames**: 16, **Batch size**: 64, **Weight Decay**: 4e-5, **Epochs**: 150
- **LR**: 1e-4, **Frames**: 16, **Batch size**: 64, **Weight Decay**: 4e-4, **Epochs**: 150
- **LR**: 1e-4, **Frames**: 20, **Batch size**: 64, **Weight Decay**: 4e-5, **Epochs**: 150

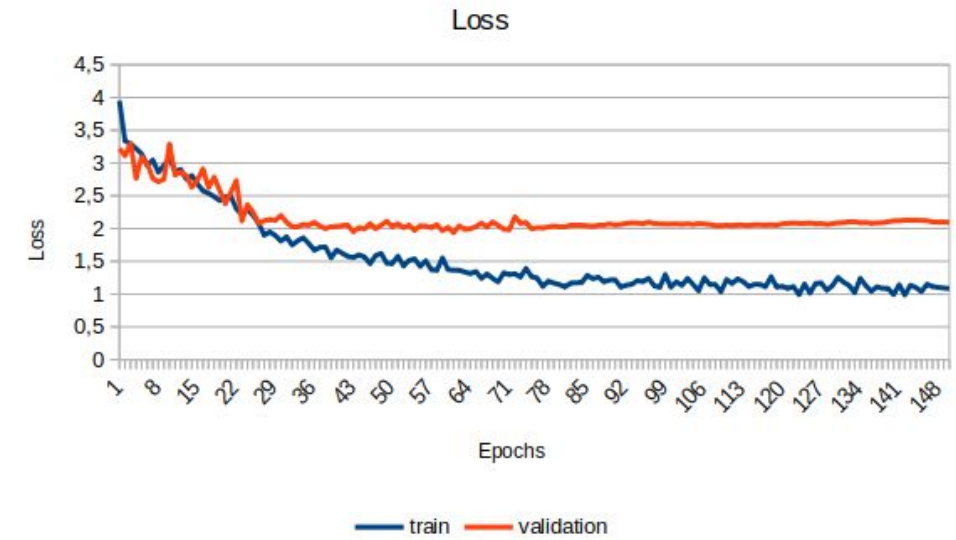
Configurations involved are:

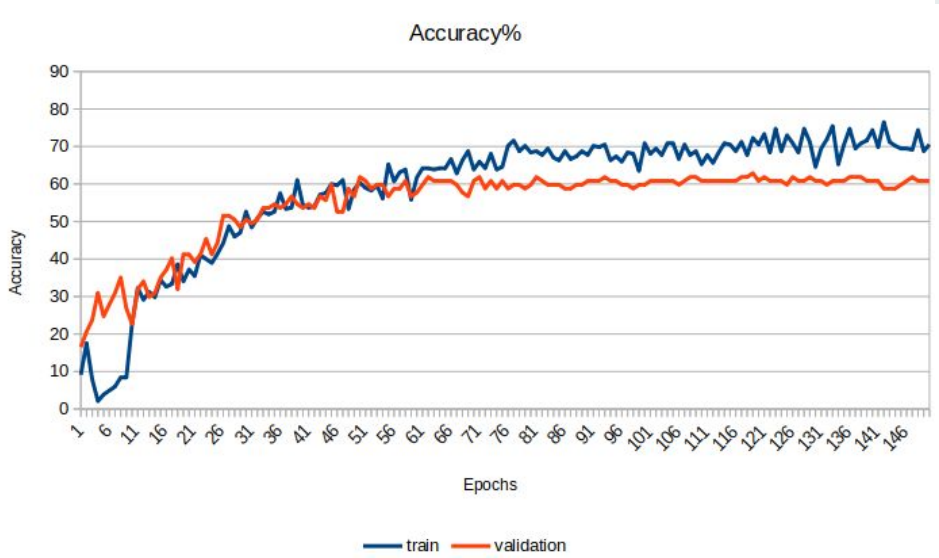
- net with only **MS**
- net with **MS** and **CAM**
- net with only **MS** in **regression** mode

Results

LR	Frames	Batch Size	Epochs	Weight Decay	MS	MS+CAM	MS_REG+CAM	MS_REG
1e-5	16	64	150	4e-5	54,72	39,62	40,57	42,45
1e-4	16	64	150	4e-4	60,38	53,77	54,72	53,77
1e-4	20	8	150	4e-5	**64,95**	61,85	59,79	60,82

Considerations





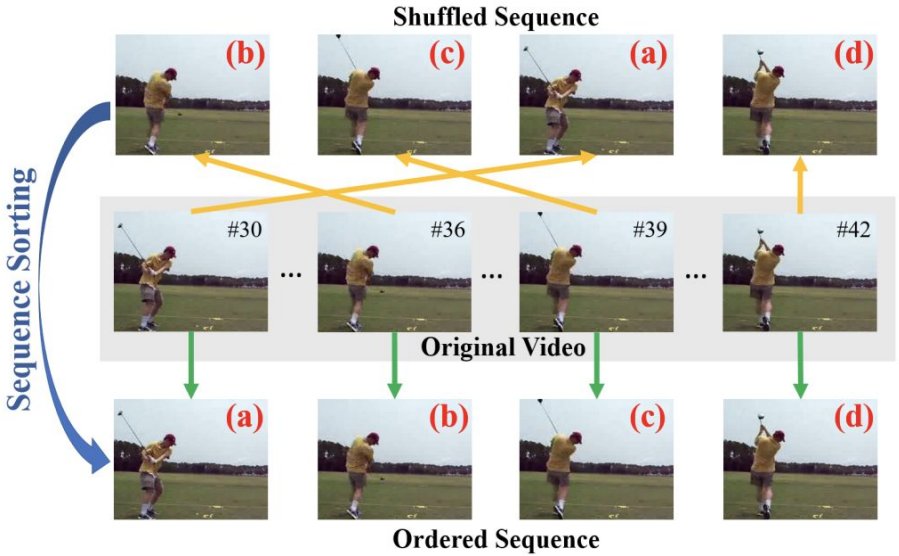
- overfitting is small
- attention mechanism with MS get worse results
- much frames improve results

Ordinal Prediction Task[*]

[*] H. Lee, J.Huang, M. Singh, M. Yang, University of California, Merced, Virginia Tech, Verisk Analytics
Unsupervised Representation Learning by Sorting Sequence

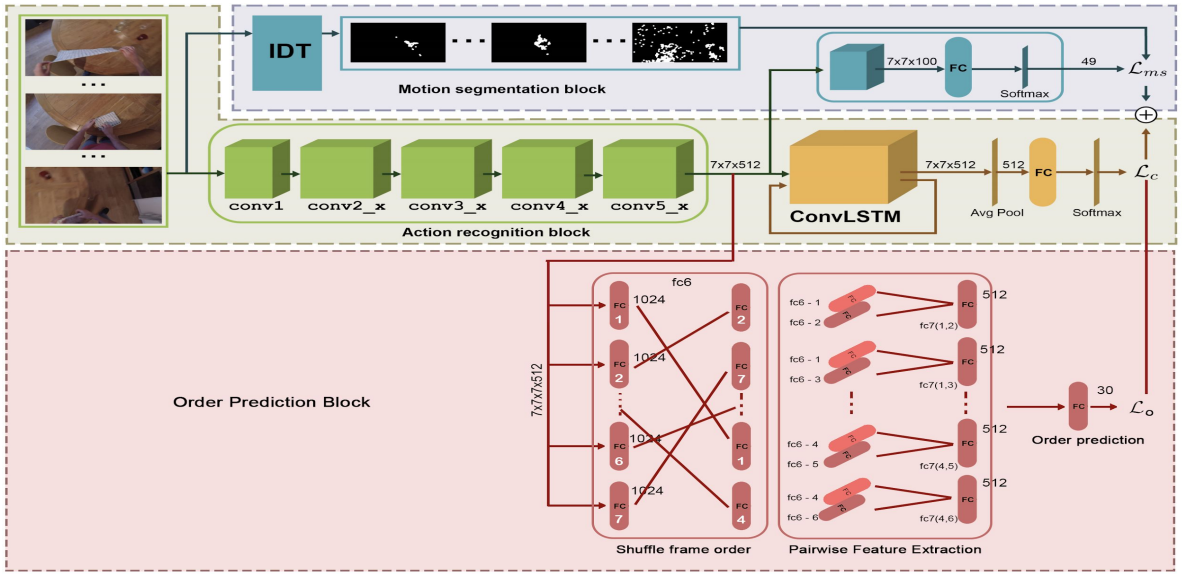
Feature Learning by Sequence Sorting

- Appearance variations and temporal coherence in videos offer rich supervisory signals for representation learning.
- Predict variations pixels is challenging
  - Multi-class classification task
  - Reduce number frames to be reordered
  - Maximum Hamming distance to select P permutations



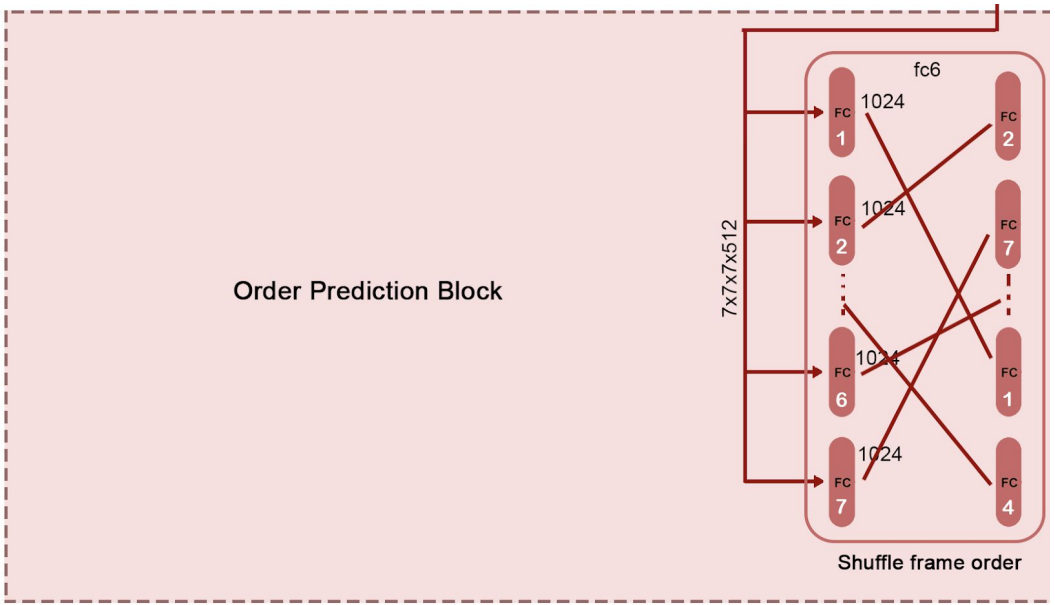
H. Lee, J.Huang, M. Singh, M. Yang, University of California, Merced, Virginia Tech, Verisk Analytics - *Unsupervised Representation Learning by Sorting Sequence*

Architecture overview



Architecture overview

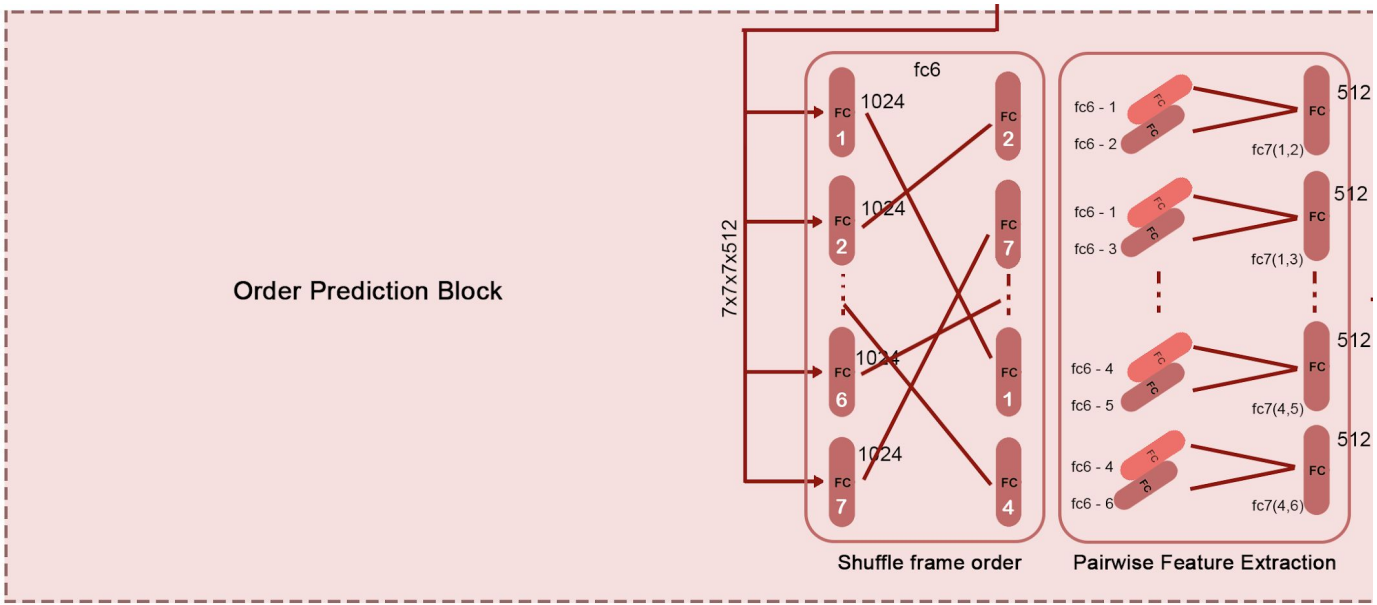
1. Shuffle feature frame order



Permutation Set	
index	permutation
64	9,4,6,8,3,2,5,1,7

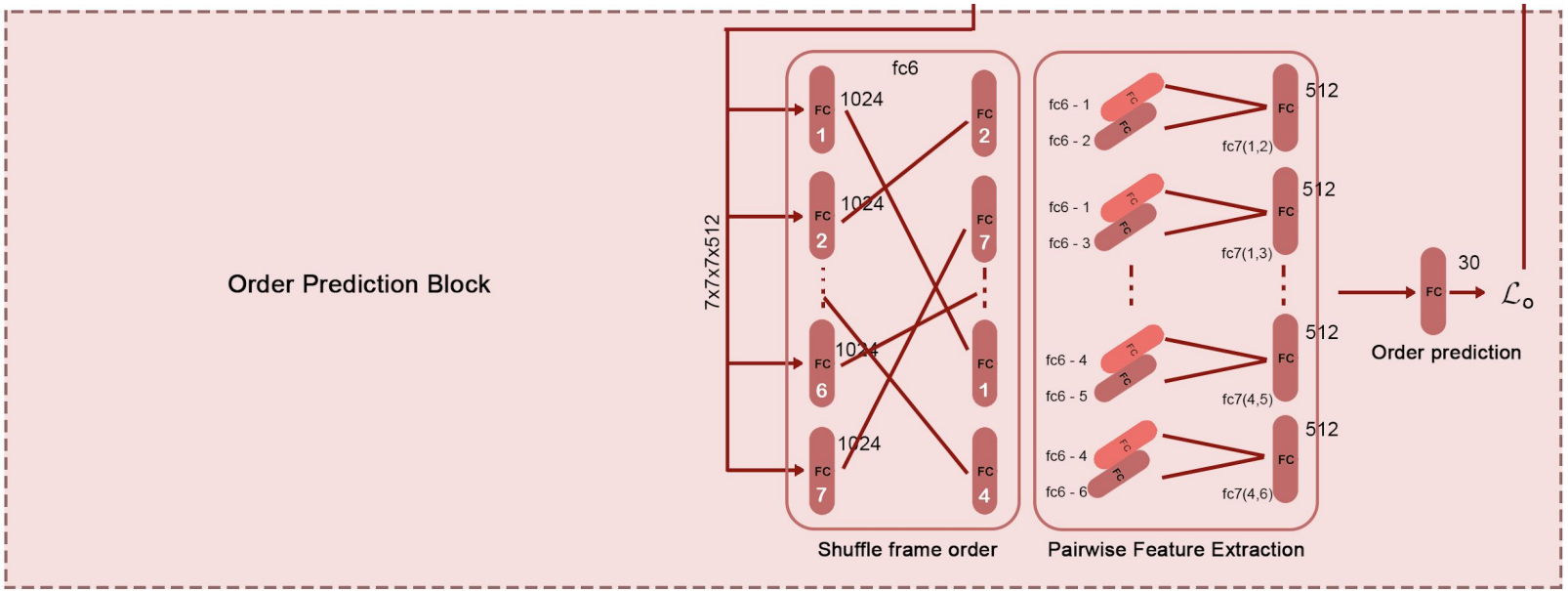
Reorder frames according to the selected random permutation

2. Pairwise Feature Extraction

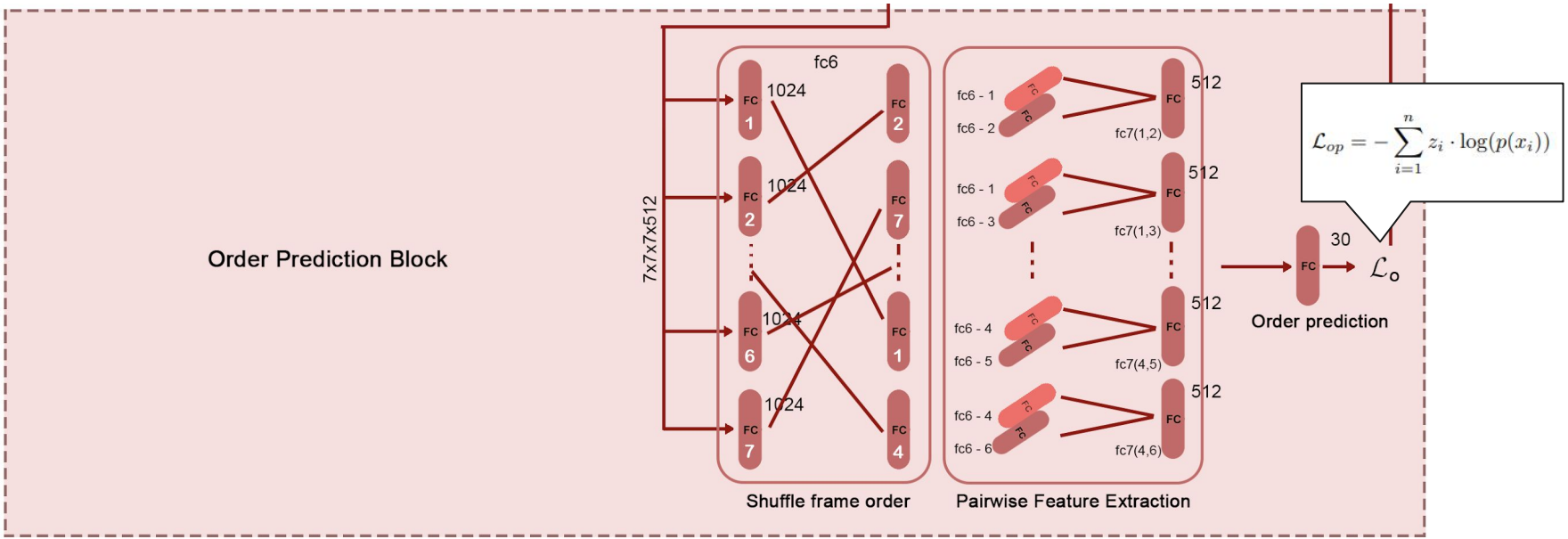


Provides informations of the relationship of a frame with another.

3. Order prediction



3. Order prediction



Results

(Sparnet$_{(7frm)}$ + Order Prediction Task))

Shuffled Frame	n!	P	SparNet + OP (%Acc)	OP (%Acc)	SparNet + OP + Attention (%Acc)	OP (%Acc)
4	24	12	**60,04**	57,05	54,71	20.3
7	5040	100	55,44	64,26	53,43	47,14
		500	56,89	56,15	55,60	26,42
		1000	54,31	44,44	53,44	3,30

(7 frames - 150 epochs - lr 10-4 - step-size [25,75])

Results

(Sparnet$_{(7frm)}$ + Order Prediction Task))

Shuffled Frame	n!	P	SparNet + OP (%Acc)	OP (%Acc)	SparNet + OP + Attention (%Acc)	OP (%Acc)
4	24	12	**60,04**	57,05	54,71	20.3
7	5040	100	55,44	64,26	53,43	47,14
		500	56,89	56,15	55,60	26,42
		1000	54,31	44,44	53,44	3,30

(7 frames - 150 epochs - lr 10-4 - step-size [25,75])

Results

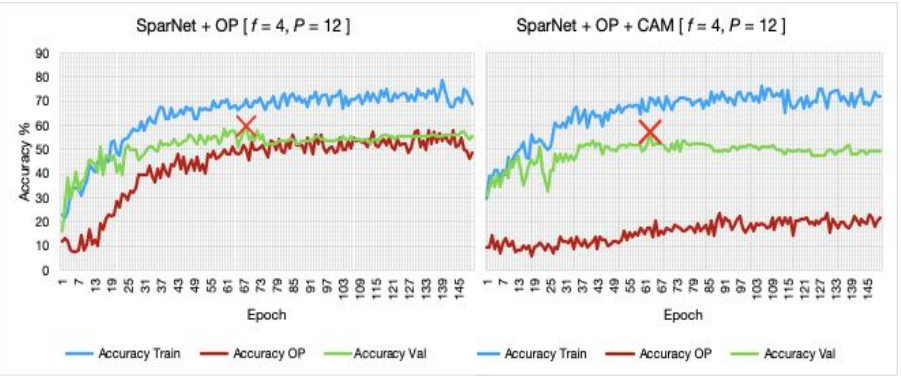
(Sparnet$_{(7frm)}$ + Order Prediction Task))

Shuffled Frame	n!	P	SparNet + OP (%Acc)	OP (%Acc)	SparNet + OP + Attention (%Acc)	OP (%Acc)
4	24	12	**60,04**	57,05	54,71	20.3
7	5040	100	55,44	64,26	53,43	47,14
		500	56,89	56,15	55,60	26,42
		1000	54,31	44,44	53,44	3,30

(7 frames - 150 epochs - lr 10-4 - step-size [25,75])

Some considerations

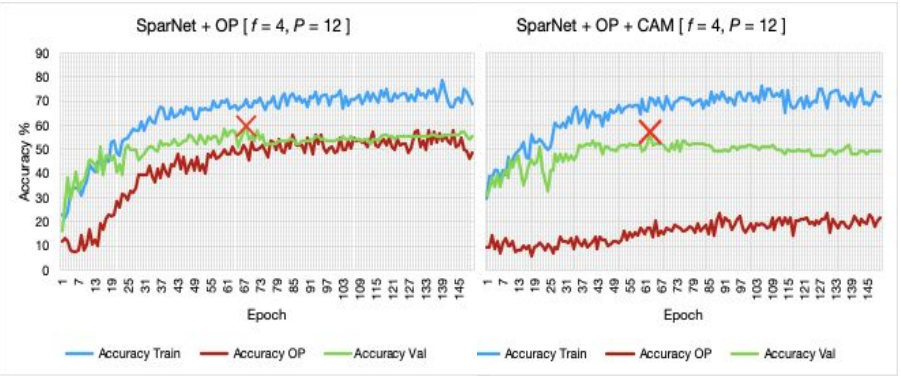
- Any improvement with CAM. The attention, similarly with MS, causes accuracy drops
- Our features are more “task-specific”



Analysis of the behaviour of SparNet with OP task (4 frames)

Some considerations

- Any improvement with CAM. The attention, similarly with MS, causes accuracy drops
- Our features are more “task-specific”



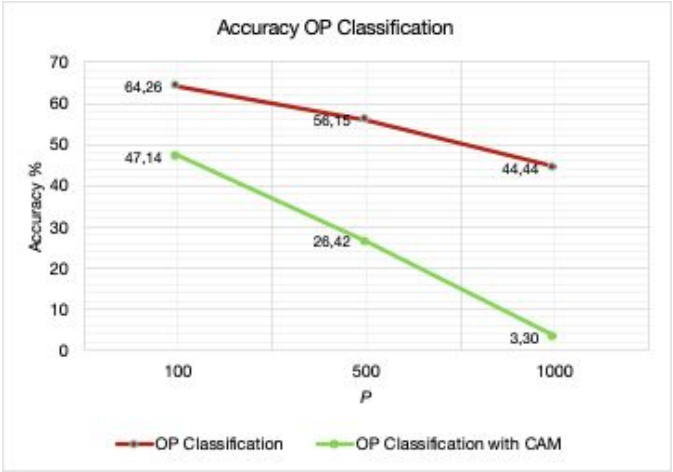
Analysis of the behaviour of SparNet with OP task (4 frames)

	%Acc	OP (%Acc)	OP + MS (%Acc)
Ego-RGB + OP	56,89	35,43	57,05
Ego-RGB + OP + Cam	52,58	20.3	20.3

Results without MS (4 frames)

Some considerations

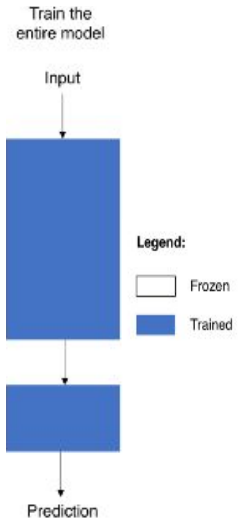
- OP classifier is producing meaningful results per-se
- The performance decreases when the task becomes more difficult



Analysis of the behaviour of OP with different permutations (7 frames)

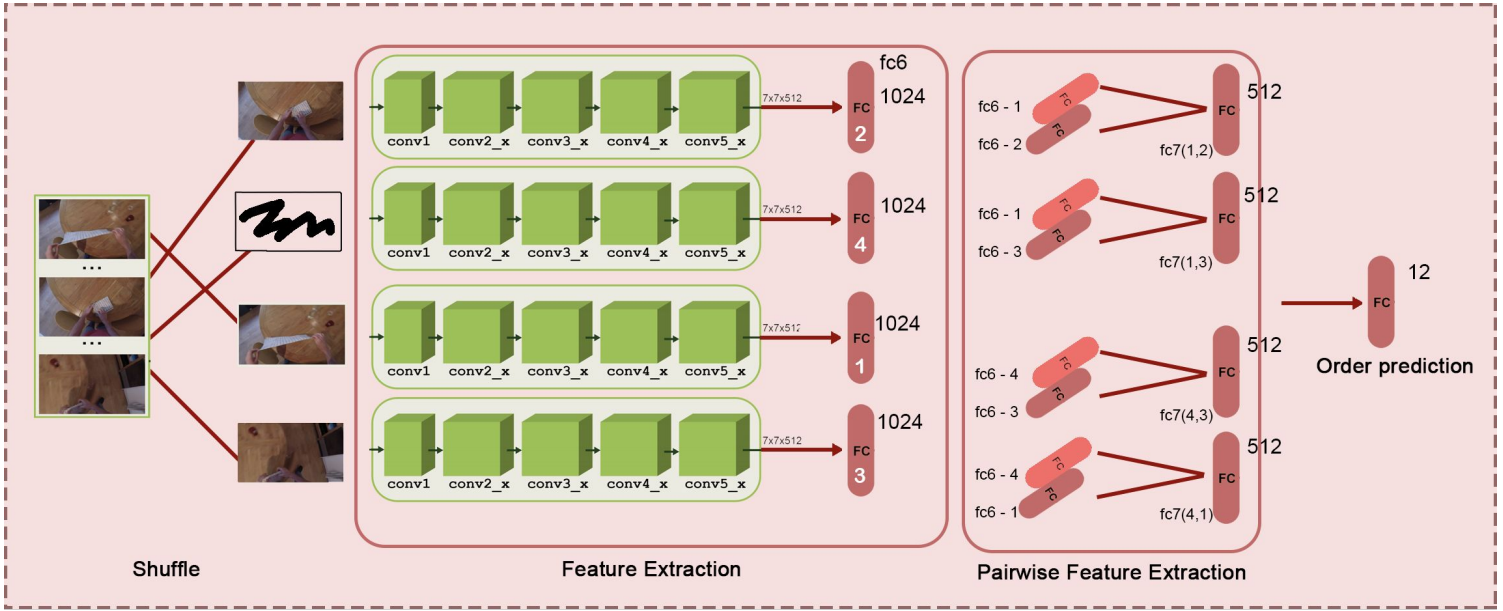
An unsupervised pre-training approach

- Preatraning-finetuning paradigm
- Train the entire network to initialize model for action recognition
- The goal is to encourage the model to reason about the motion and appearance of the objects, and thus learn the temporal structure of videos



https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751

Architecture overview



Order Prediction Network

Results

(Transfer Learning)

	Pre-training on OPN	SparNet (7 frames) + Attention	SparNet (7 frames)
Accuracy %	20,35	38,79	45,68

(300 epochs - lr 10-3 - step-size [50,100,150])

(7 frames - 150 epochs - lr 10-4 - step-size [25,75])

- No improvement
- **Reasons**:
  - Small dataset (457 samples)
  - Computational power
- Further insights are needed

THANK YOU
FOR ATTENTION

POLITECNICO
DI TORINO