UCF
Computational Imaging Lab. (CIL)
Department of Computer Science

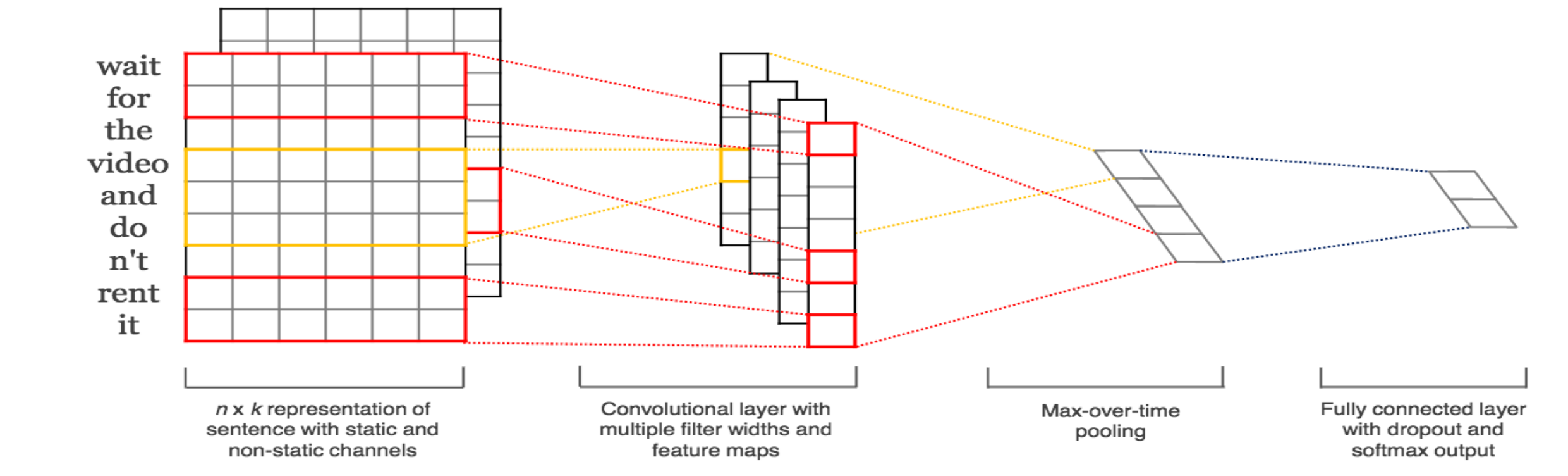
A Temporal Sequence Learning for Action Recognition and Prediction

Sangwoo Cho, Hassan Foroosh (swcho@knights.ucf.edu, foroosh@cs.ucf.edu) Dept. of Computer Science, University of Central Florida

WACV 2018

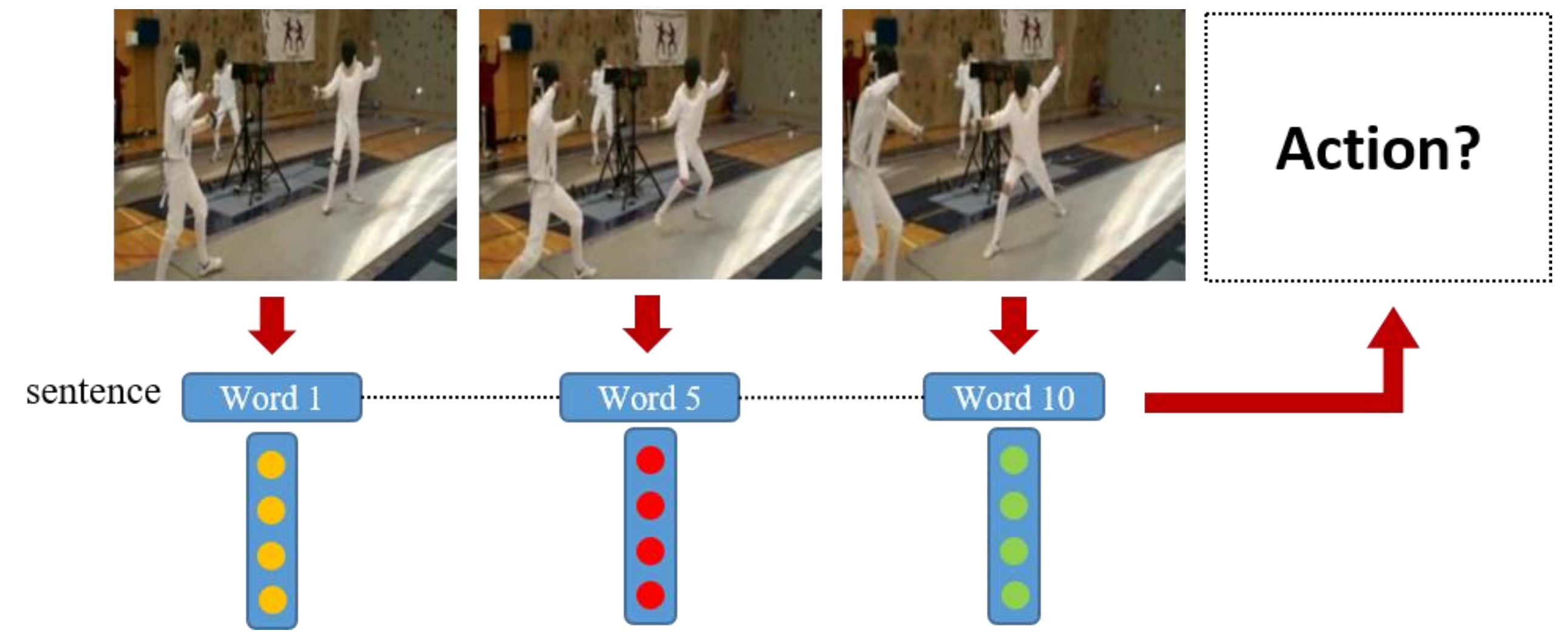
Motivation

Sentence = a sequence of words


wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

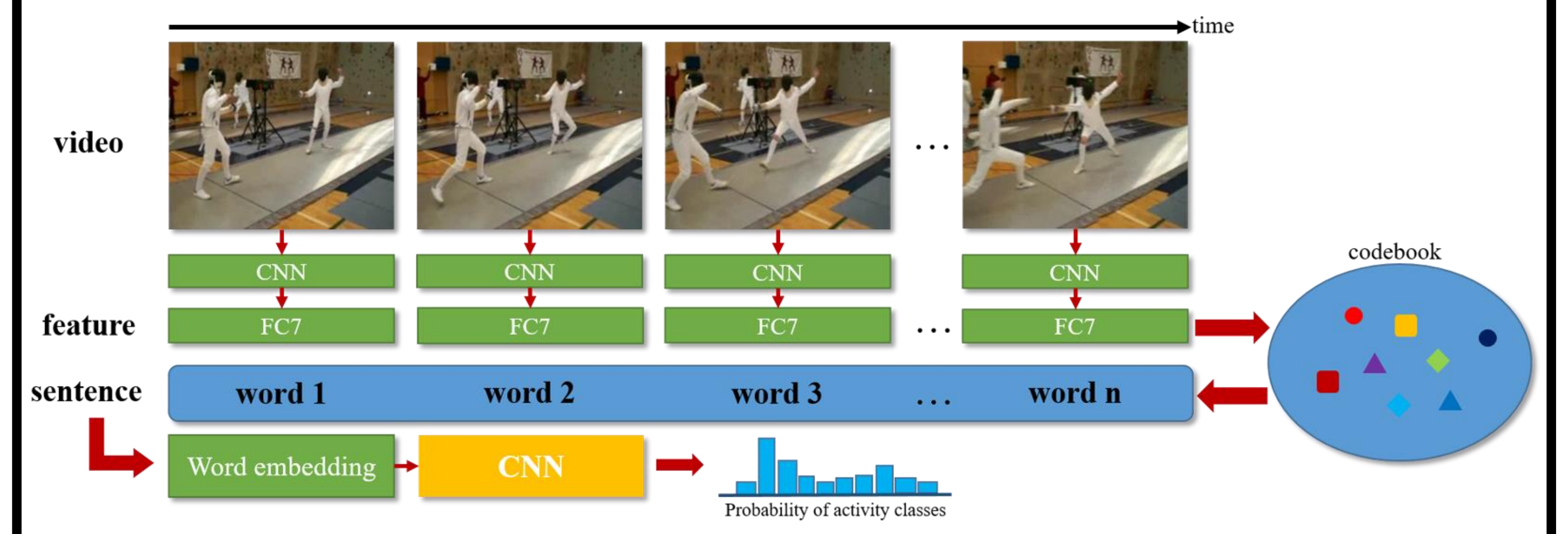
Yoon Kim, "Convolutional Neural Networks for Sentence Classification", 2014

Video = a sequence of frames



Q) How to represent each frame and train them?

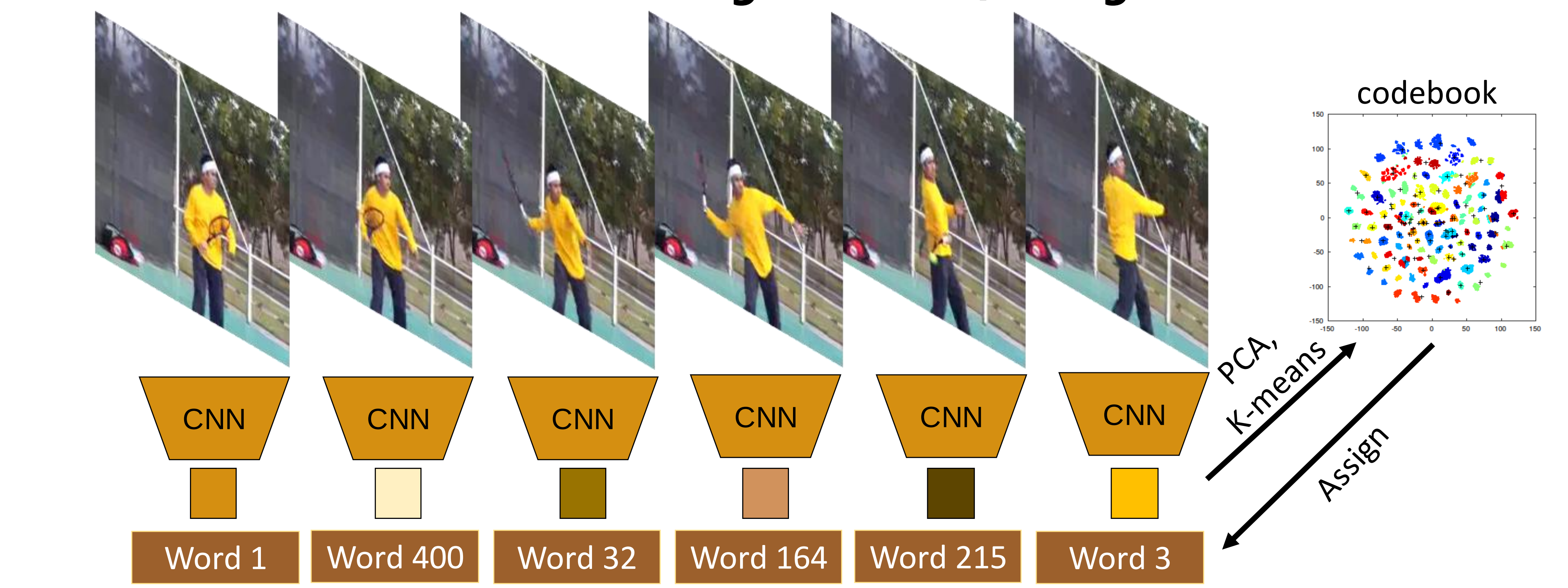
System overview

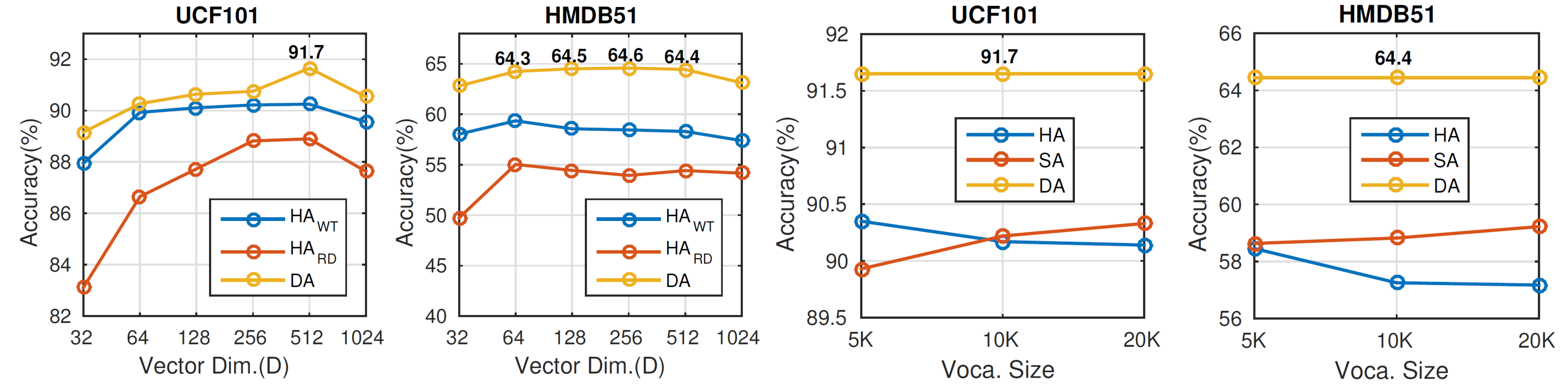


- *Feature extraction* using CNN (VGG-16)
- BoW based codebook generation / Assignment
- Two stream data fusion with optimal data ratio
- Sequence training using *Temporal CNNs*

Method

BoW based Codebook generation / Assignment

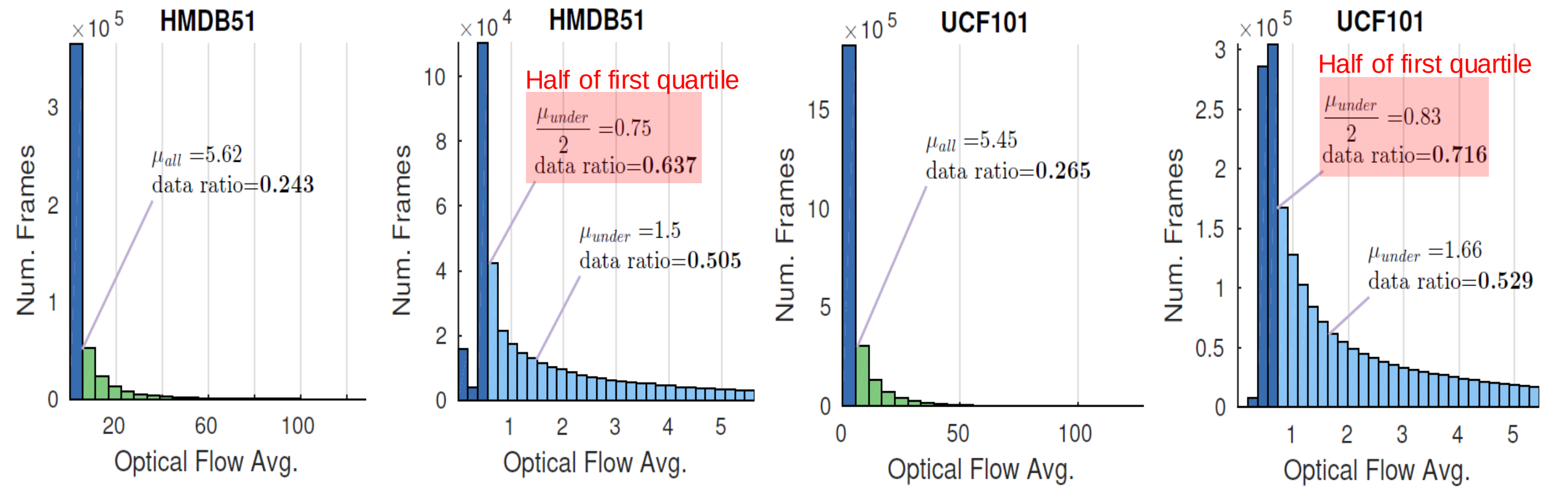




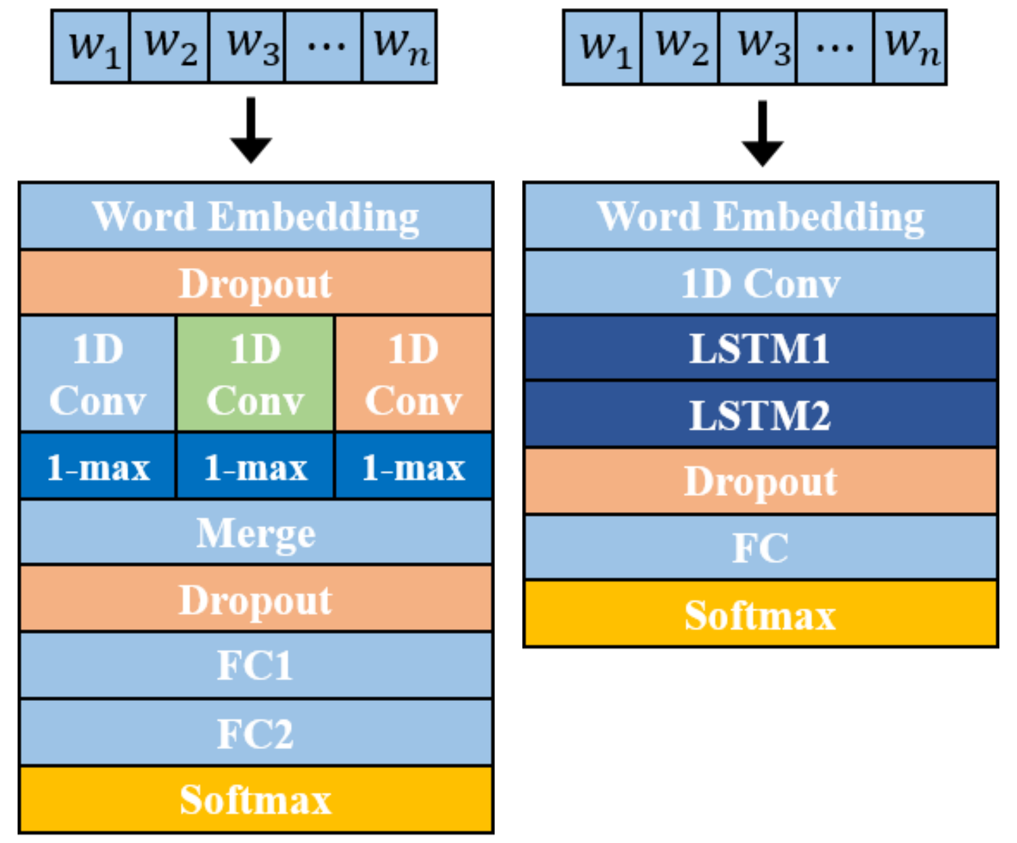
Optimal fusion ratio

Feature = $L_{RGB} \times (1-r)$ | $L_{FLOW} \times r$

HMDB51	$r = 0.5$	$r = 0.625$	$r = 0.75$
512 dim.	64.8	**66.4**	65.1
UCF101	$r = 0.5$	$r = 0.625$	$r = 0.75$
512 dim.	91.5	91.8	**92.7**



Sequence learning



✓ Training time (UCF101 / HMDB51)
- 51min / 10min 101min / 67min
✓ Feature extraction time
- ~12hrs / ~5hrs

Experimental results

VGG-16 baseline performance

	UCF101	HMDB51
Spatial	81.8	44.8
Temporal	84.9	55.0
Two-stream	90.1	61.4

Action recognition performance

HMDB51		UCF101	
iDT+FV	57.2	iDT+FV	85.9
Two stream	59.4	Two stream	88.0
TDD+FV	63.2	TDD+FV	90.3
Transformation	62.0	Transformation	92.4
KVMF	63.3	KVMF	93.1
Fusion net	65.4	Fusion net	92.5
Ours(C-LSTM)	62.4	Ours(C-LSTM)	90.9
Ours(T-CNN) HA*	61.9	Ours(T-CNN) HA*	90.5
Ours(T-CNN) HA†	62.3	Ours(T-CNN) HA†	91.1
Ours(T-CNN) SA‡	62.8	Ours(T-CNN) SA‡	91.3
Ours(T-CNN) DA	**66.3**	Ours(T-CNN) DA	**92.5**

*: HA with random weights, 5k codebook(assignment only), 512 dim. ➔ ***only sequence number!***
†: HA with 5k codebook weights, 512 dim.
‡: SA with 20k codebook weights, 512 dim.

Action prediction performance

UCF101	0-10%	0-20%	0-30%	0-40%	0-50%	0-60%	0-70%	0-80%
MOS		35.0		37.1		39.4		40.3
SMMED		40.6		40.6		40.6		40.6
Fusion	**82.8**	85.5	87.5	88.8	89.2	90.4	90.7	91.0
Ours	82.2	**86.7**	**88.5**	**89.5**	**90.1**	**91.0**	**91.5**	**91.9**
HMDB51	**0-10%**	**0-20%**	**0-30%**	**0-40%**	**0-50%**	**0-60%**	**0-70%**	**0-80%**
Fusion	**44.8**	51.5	54.5	58.0	61.0	62.9	64.9	65.2
Ours	38.8	**51.6**	**57.6**	**60.5**	**62.9**	**64.6**	**65.6**	**66.2**