LEHIGH UNIVERSITY

ReHAR: Robust and Efficient Human Activity Recognition

Xin Li, Mooi Choo Chuah

WACV18'

OUTLINE

- Motivation
- The state-of-the-art scheme
- Our solution
- Evaluations
- Why does it work
- References

OUTLINE

- Motivation
- The state-of-the-art scheme
- Our solution
- Evaluations
- Why does it work
- References

Motivation

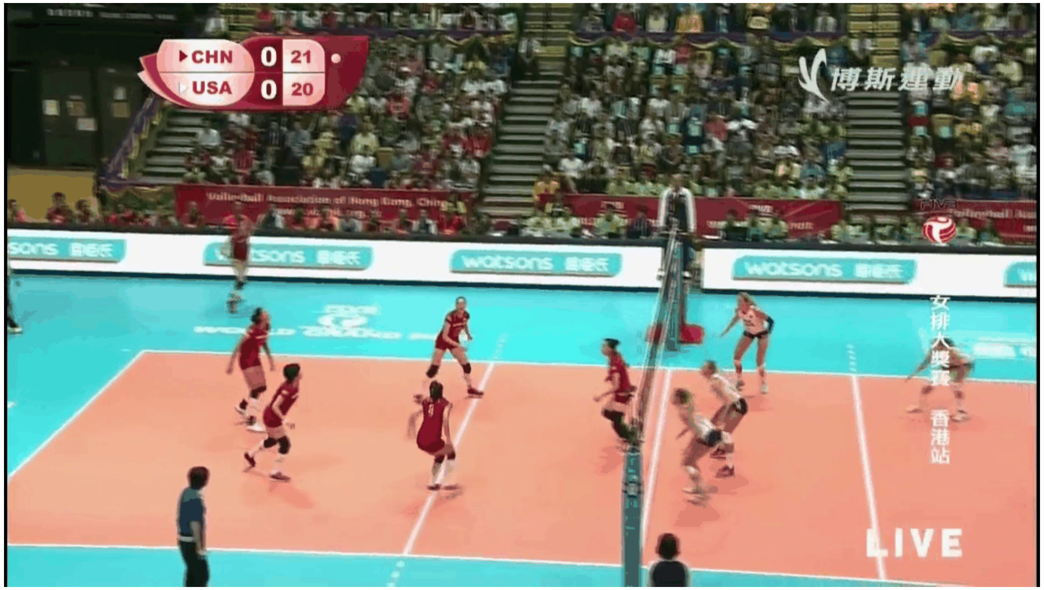
Large amount of Videos

Motivation

Public Safety

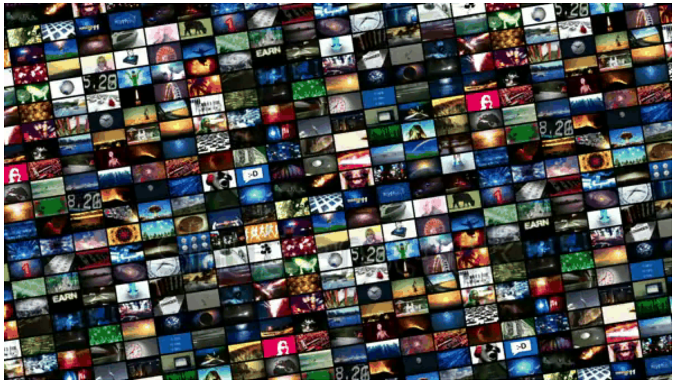
Motivation

Key events in sport videos



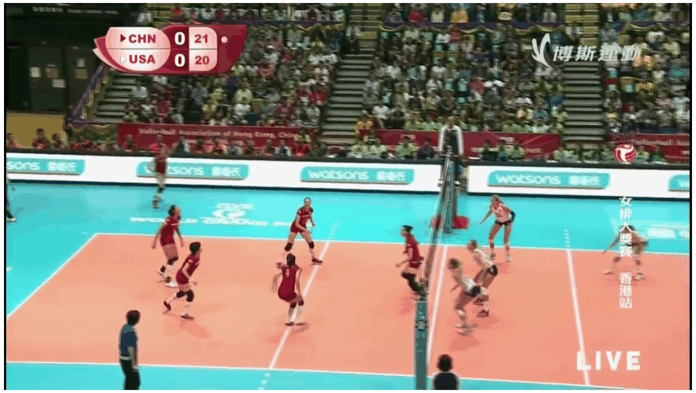
Motivation

Search among Videos



Public Safety

Game highlights



An efficient scheme for identifying activities is critically important.

OUTLINE

- Motivation
- The state-of-the-art scheme
- Our solution
- Evaluations
- Why does it work
- References

Building Block

Long Short Term Memory Network (LSTM)

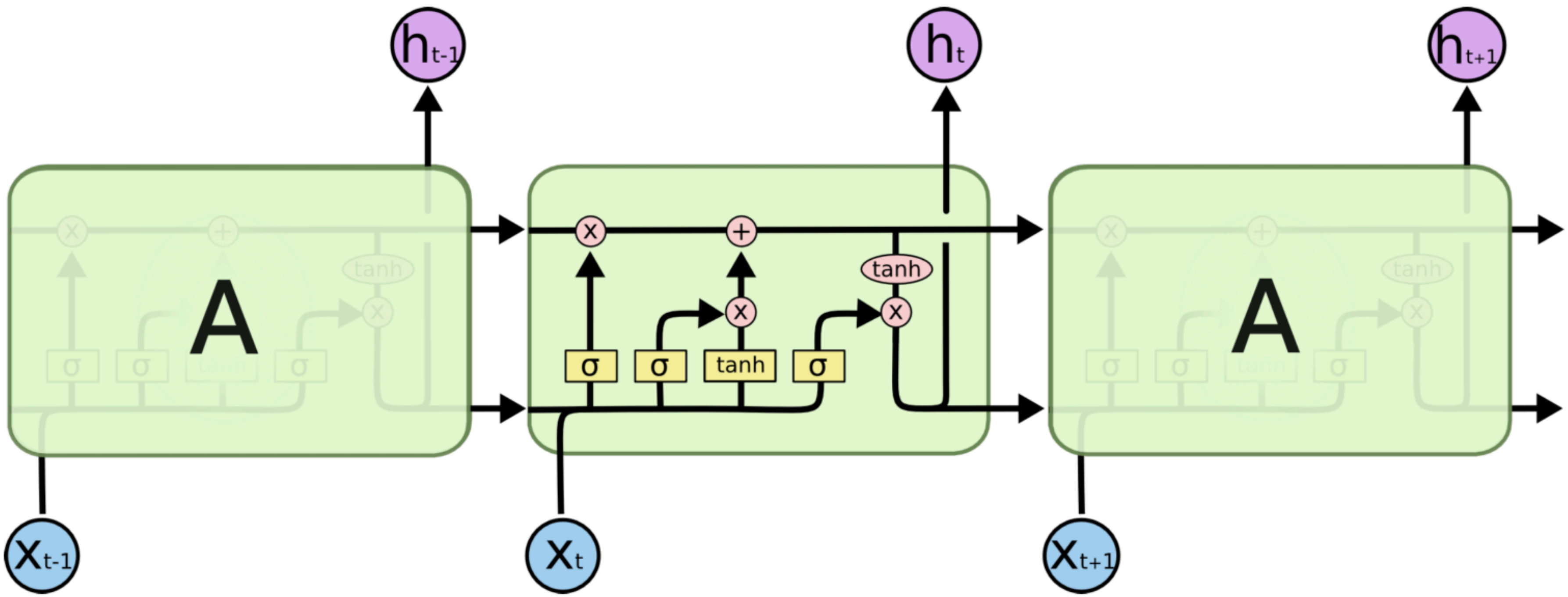


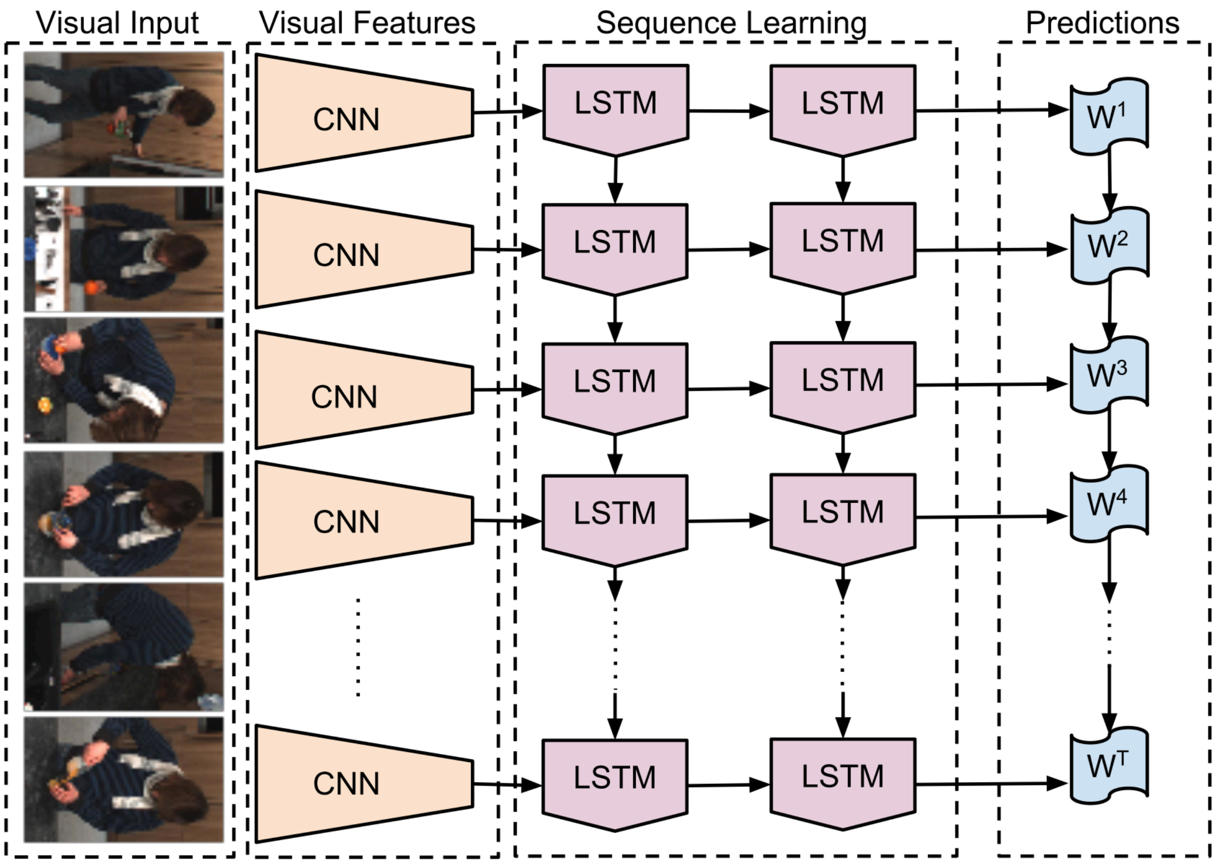
Figure from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Existing Work

[1] Jeff Donahue, et al.

Long-term recurrent convolutional networks for visual recognition and description

CVPR. 2015

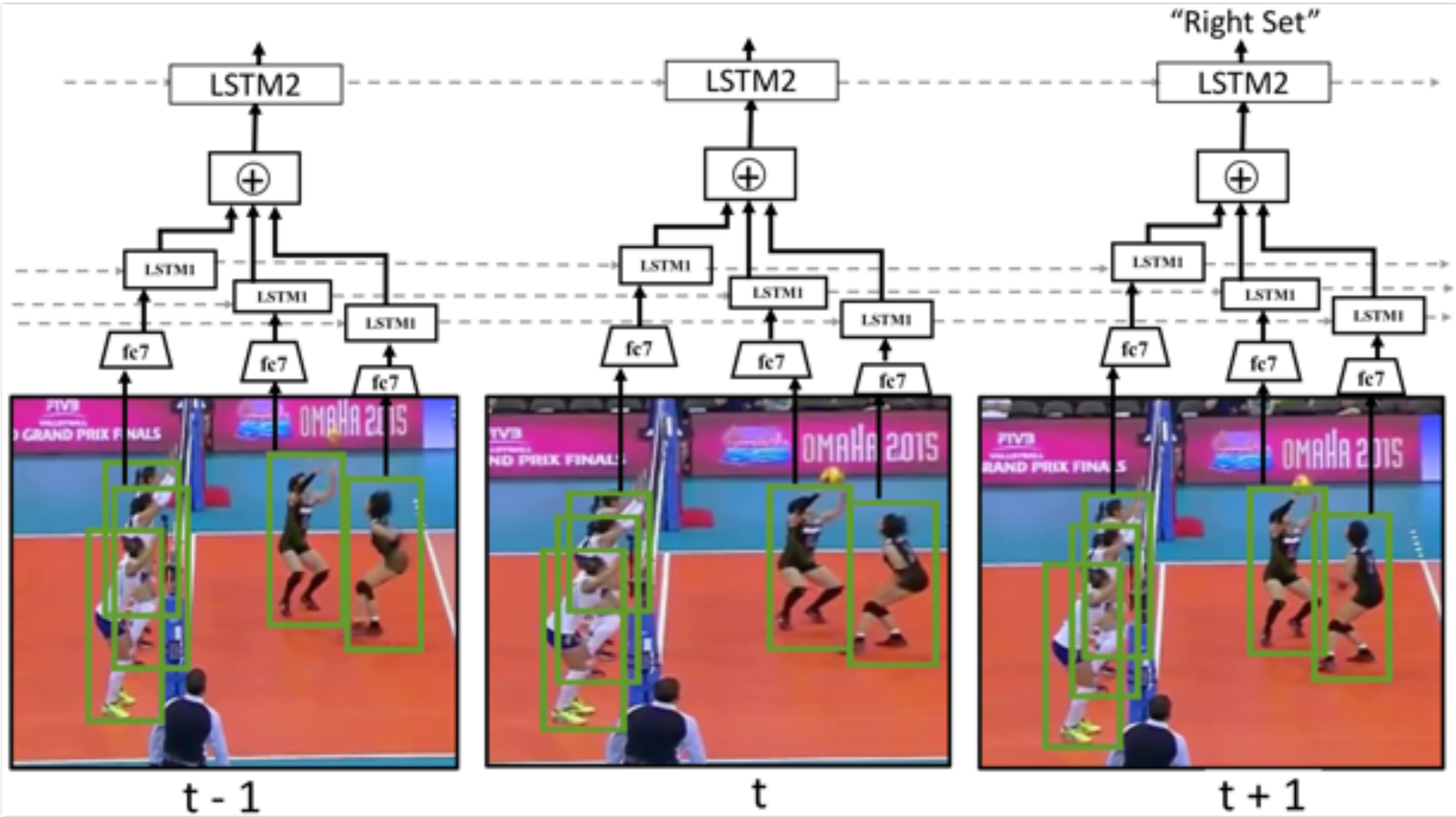


Existing Work

[2] Ibrahim Moustafa, et al.

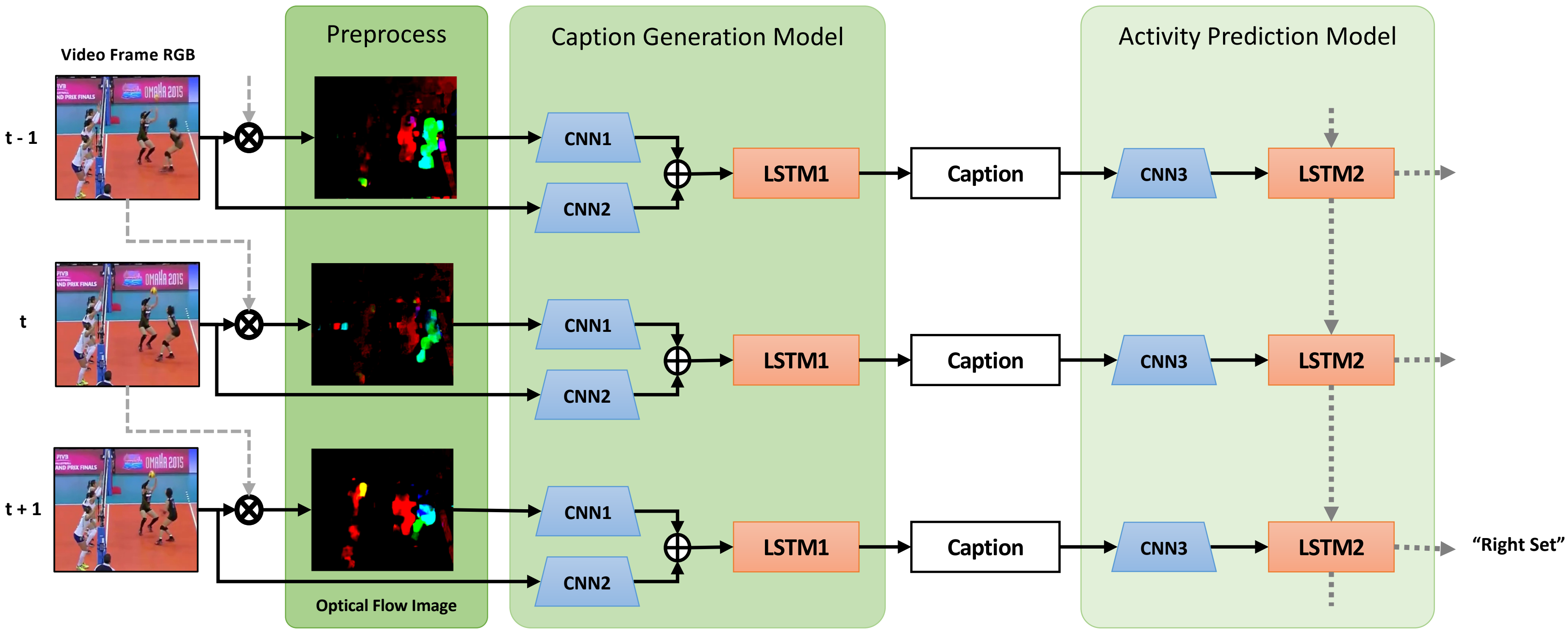
A Hierarchical Deep Temporal Model for Group Activity Recognition

CVPR. 2016



Existing Work

[3] Xin Li, Mooi Choo Chuah
SBGAR: Semantics Based Group Activity Recognition
ICCV. 2017

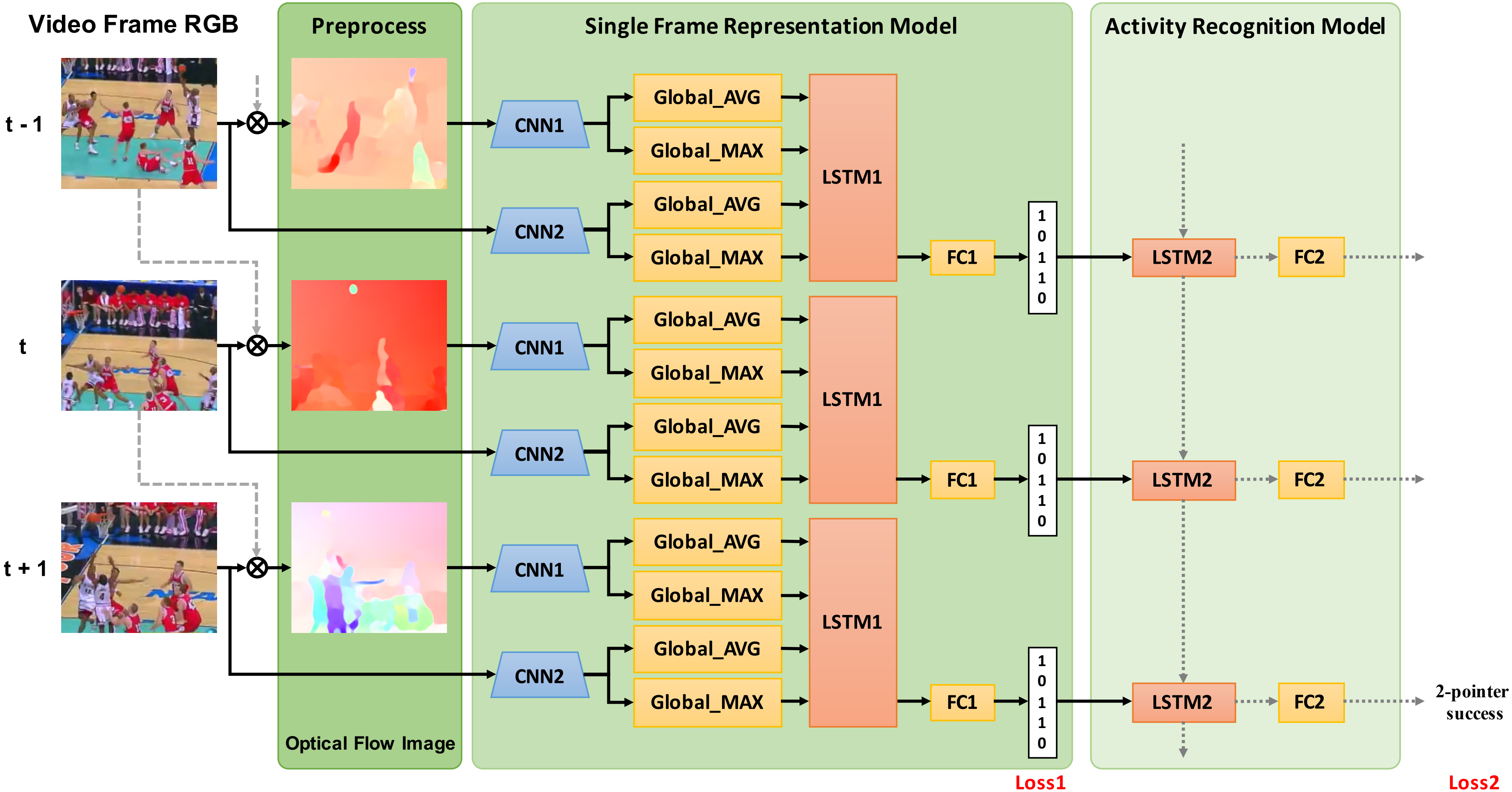


OUTLINE

- Motivation
- The state-of-the-art scheme
- Our solution
- Evaluations
- Why does it work
- References

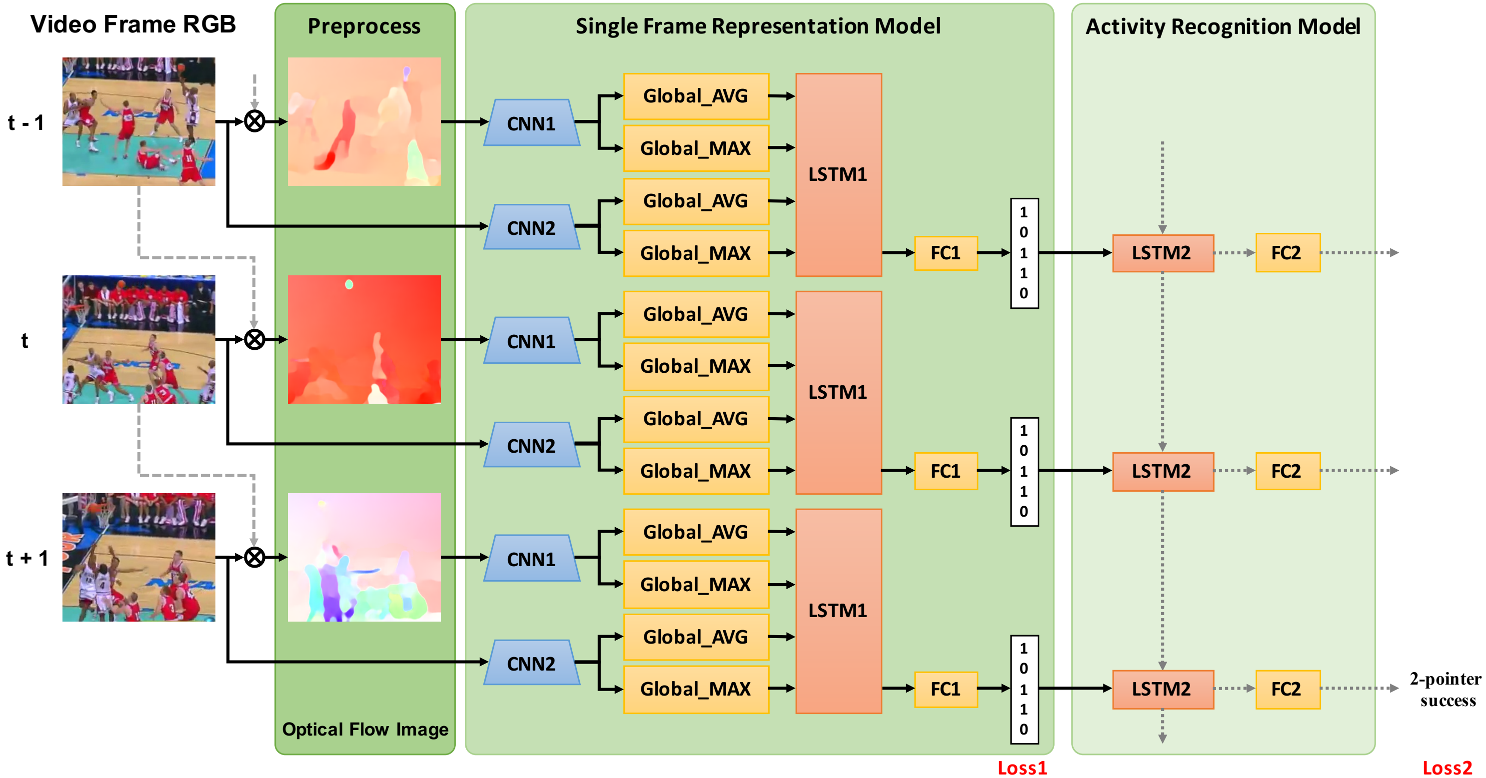
Activity Recognition

Our Solution



Activity Recognition

Our Solution



$$Loss = \left(\sum_{t=1}^{time\_step} loss_{1,t} \right) + \lambda * loss_2$$

OUTLINE

- Motivation
- The state-of-the-art scheme
- Our solution
- Evaluations
- Why does it work
- References

Dataset1: NCAA Basketball Dataset

NCAA Basketball dataset:

11436 training videos
856 validation videos
2256 testing videos

Event	No. of videos Train (Test)
3-point succ.	895 (188)
3-point fail.	1934 (401)
free-throw succ.	552 (94)
free-throw fail.	344 (41)
layup succ.	1212 (233)
layup fail.	1286 (254)
2-point succ.	1039 (148)
2-point fail.	2014 (421)
slam dunk succ.	286 (54)
slam dunk fail.	47 (5)
steal	1827 (417)

Test Result using NCAA Basketball Dataset

	3point S.	3point F.	throw S.	throw F.	layup S.	layup F.	2point S.	2point F.	dunk S.	dunk F.	steal	Mean
IDT[4]	0.370	0.501	0.778	0.365	0.283	0.278	0.136	0.303	0.197	0.004	0.555	0.343
IDT[4] player	0.428	0.481	0.703	0.623	0.300	0.311	0.233	0.285	0.171	0.010	0.473	0.365
C3D[5]	0.117	0.282	0.642	0.319	0.195	0.185	0.078	0.254	0.047	0.004	0.303	0.221
MIL[6]	0.237	0.335	0.597	0.318	0.257	0.247	0.224	0.299	0.112	0.005	0.843	0.316
LRCN[7]	0.462	0.564	0.876	0.584	0.463	0.386	0.257	0.378	0.285	**0.027**	0.876	0.469
Atten. no track[8]	0.583	0.668	0.892	0.671	0.489	0.426	0.281	0.442	0.210	0.006	0.886	0.505
Atten. track[8]	0.600	0.738	0.882	0.516	0.500	**0.445**	0.341	0.471	**0.291**	0.004	0.893	0.516
Ours	**0.753**	**0.766**	**0.933**	**0.857**	**0.613**	0.435	**0.405**	**0.542**	0.232	0.007	**0.940**	**0.589**

[4] Heng Wang, Alexander Kla¨ser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In CVPR, 2011.

[5] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. CoRR, abs/1412.0767, 2(7):8, 2014.

[6] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In Advances in neural information processing systems, pages 577–584, 2003.

[7] Jeffrey Donahue,Lisa Anne Hendricks,Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In ICCV, pages 2625–2634, 2015.

[8] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. ICCV, 2016.

Test Result using NCAA Basketball Dataset

	3point S.	3point F.	throw S.	throw F.	layup S.	layup F.	2point S.	2point F.	dunk S.	dunk F.	steal
3point S.	61.17	11.17	0.53	0.00	2.13	1.60	18.09	4.79	0.00	0.00	0.53
3point F.	1.75	72.07	0.00	0.00	0.00	1.00	1.00	23.19	0.00	0.00	1.00
throw S.	1.06	0.00	87.23	6.38	3.19	0.00	1.06	0.00	0.00	0.00	1.06
throw F.	0.00	4.88	17.07	75.61	0.00	0.00	0.00	0.00	0.00	0.00	2.44
layup S.	2.58	1.29	0.43	0.00	59.66	12.02	15.88	6.01	1.72	0.00	0.43
layup F.	0.00	3.94	0.00	0.00	8.66	47.64	1.57	35.83	0.00	0.00	2.36
2point S.	12.16	4.05	0.68	0.00	31.08	6.08	36.49	7.43	0.00	0.00	2.03
2point F.	1.66	16.86	0.00	0.24	1.19	17.10	0.95	58.43	0.00	0.00	3.56
dunk S.	0.00	0.00	0.00	1.85	53.70	24.07	9.26	3.70	5.56	0.00	1.85
dunk F.	0.00	0.00	0.00	0.00	0.00	60.00	0.00	20.00	0.00	0.00	20.00
steal	0.00	4.32	0.00	0.24	1.92	5.04	0.00	6.24	0.00	0.00	82.25

Test Result using NCAA Basketball Dataset

Group 10 shooting-related actions (except "steal") into 2 categories (success or failure)

	Number of testing videos					
	3-point	**free-throw**	**layup**	**2-point**	**slam dunk**	**In total**
Success	188	94	233	148	54	717
Failure	401	41	254	421	5	1122

88% testing samples are correctly labeled into "Success" or "Failure" categories.

Dataset2: UCF Sports Action Dataset

UCF Sports dataset:

103 training videos

47 testing videos

10 different sports categories

- Diving
- Golf
- Kicking
- Lifting
- Riding
- Run
- SkateBoarding
- Swing-Bench
- Swing-Side
- Walk

Test Result UCF Sports Action Dataset

	Diving	Golf	Kicking	Lifting	Riding	Run	SkateB	Swing	SwingB	Walk	mAP
Gkioxari et al. [9]	0.758	0.693	0.546	0.991	0.896	0.549	0.298	0.887	0.745	0.447	0.681
Weinzaepfel et al. [10]	0.607	0.776	0.653	1.000	0.995	0.526	0.471	0.889	0.629	0.644	0.719
Peng et al. [11]	0.961	0.805	0.735	0.992	0.976	0.824	0.574	0.836	0.985	0.760	0.845
Hou et al. [12]	0.844	0.908	0.865	0.998	1.000	**0.837**	**0.687**	0.658	0.996	0.878	0.867
Ours	**1.000**	**0.955**	**1.000**	**1.000**	**1.000**	0.806	0.626	**1.000**	**1.000**	**0.888**	**0.928**

[9] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 759–768, 2015.

[10] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action local- ization. In Proceedings of the IEEE international conference on computer vision, pages 3164–3172, 2015.

[11] Xiaojiang Peng and Cordelia Schmid. Multi-region two- stream r-cnn for action detection. In European Conference on Computer Vision, pages 744–759. Springer, 2016.

[12] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolu- tional neural network (t-cnn) for action detection in videos. arXiv preprint arXiv:1703.10664, 2017.

Test Result using UCF Sports Action Dataset

	Driving	Golf	Kicking	Lifting	Riding	Run	SkateB.	Swing	SwingB.	Walk
Driving	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Golf	0.00	83.33	16.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kicking	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lifting	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
Riding	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.00	0.00	0.00	75.00	25.00	0.00	0.00	0.00
SkateB.	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
Swing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
SwingB.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
Walk	0.00	14.29	0.00	0.00	0.00	0.00	28.57	0.00	0.00	57.14

Computation Time

CNN base net	Time on 10 Frames (ms)	Time on 24 Frames (ms)
VGG16	103.65	239.04
InceptionV3	78.40	192.02

SBGAR [3] model using

InceptionV3 as feature extractor and

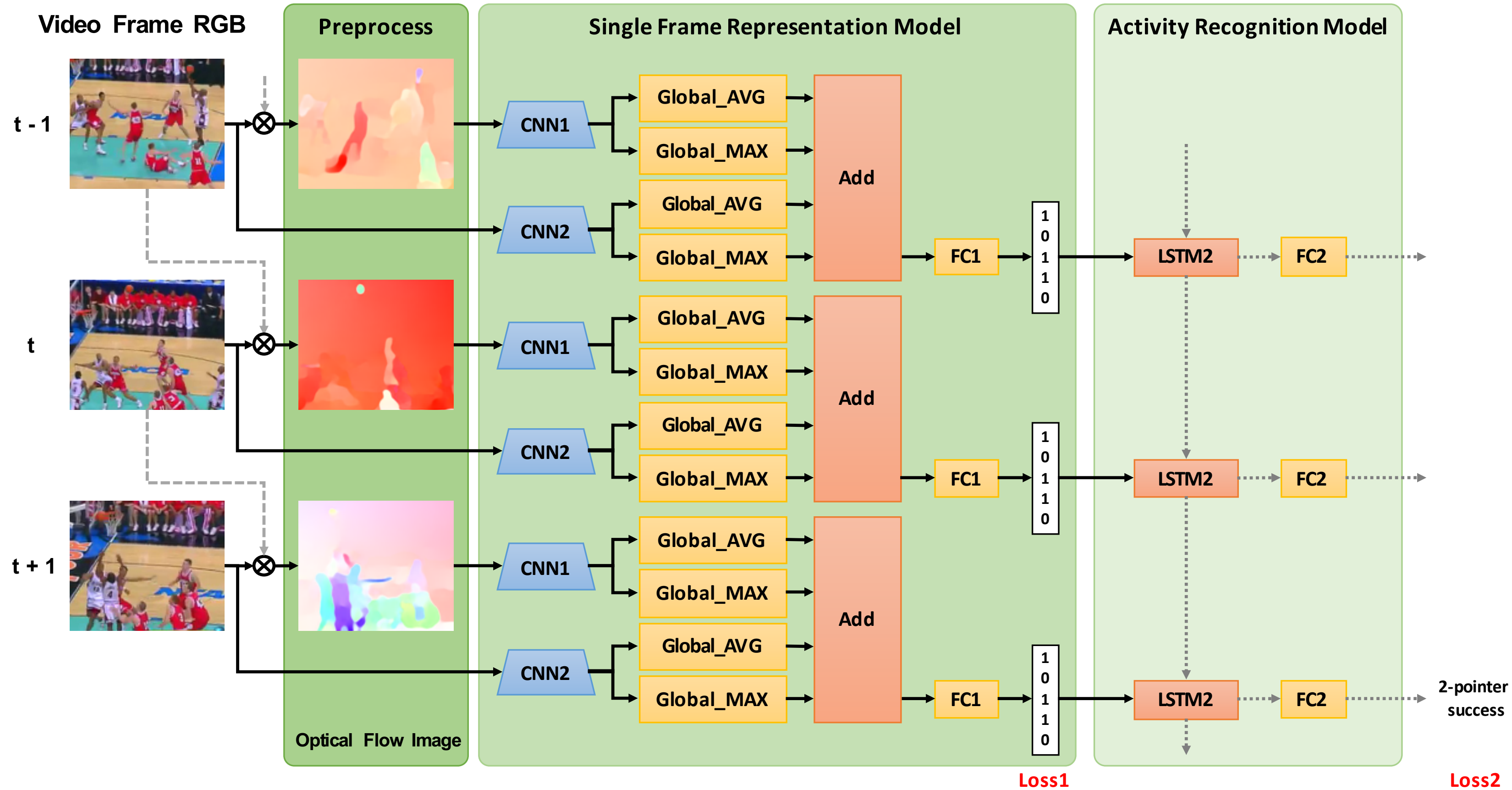
10 input frames was

108.53 ms.

OUTLINE

- Motivation
- The state-of-the-art scheme
- Our solution
- Evaluations
- Why does it work
- References

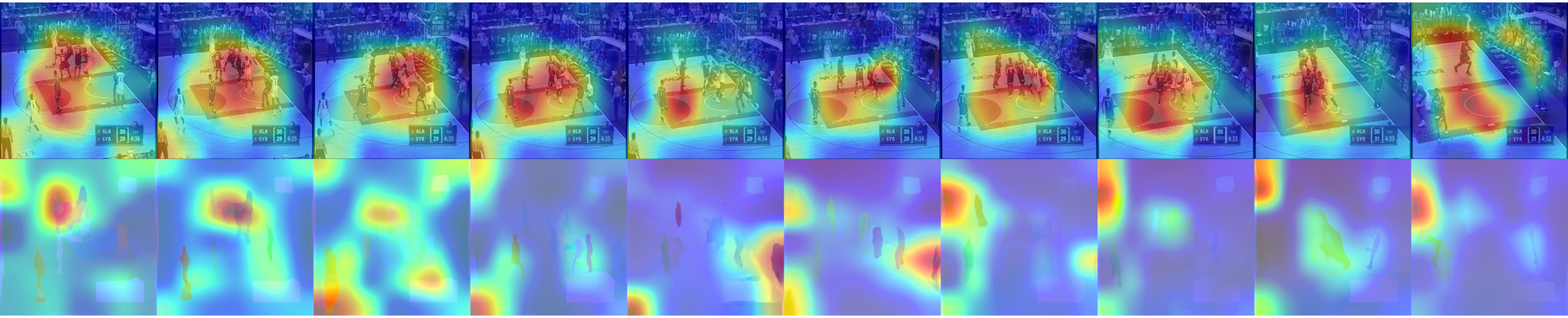
Why does our model work?



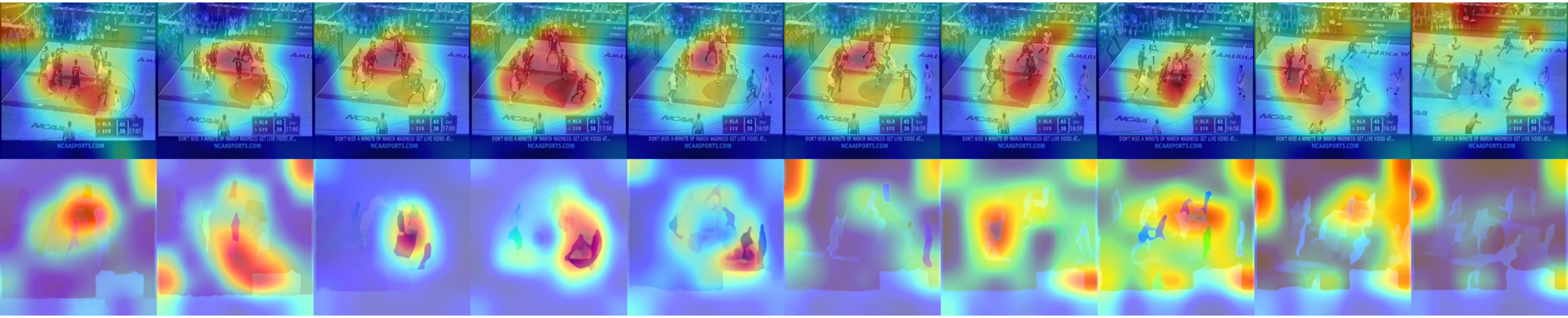
On UCF Sports Dataset

Our Model
0.928

Why does our model work?

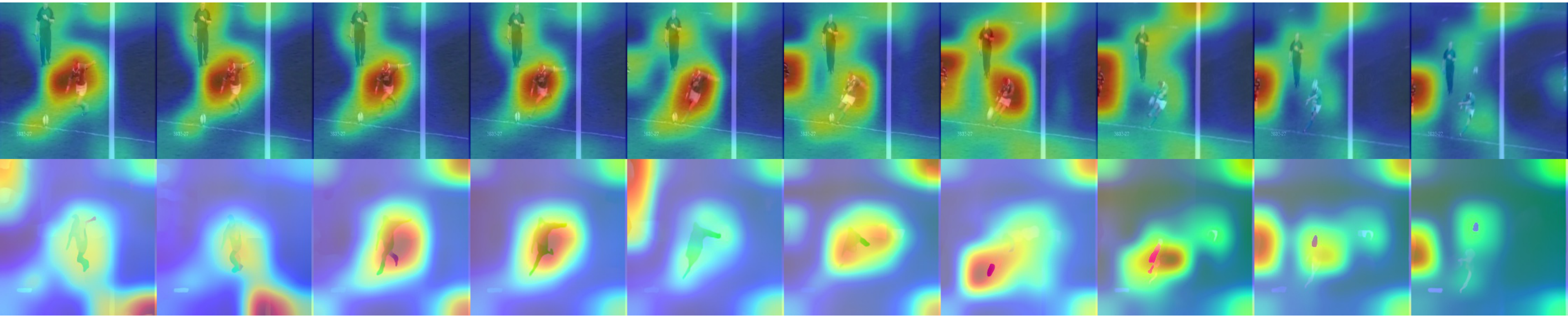


(a) Correctly predict an “other 2-pointer success” event on Basketball Dataset.

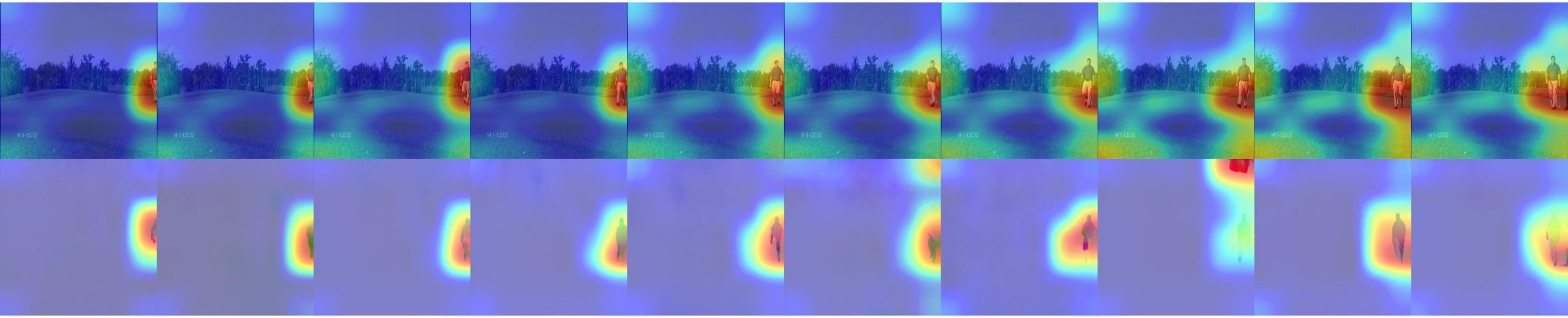


(b) Correctly predict a “Steal Success” event on Basketball Dataset.

Why does our model work?



(c) Correctly predict a “Kicking” event on UCF Sports Action Dataset.



(d) Incorrectly predict a “Walking” event as “Golf” on UCF Sports Action Dataset.

Reference

- [1] Jeff Donahue, et al., “Long-term recurrent convolutional networks for visual recognition and description,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
- [2] Ibrahim, Moustafa, et al., "A Hierarchical Deep Temporal Model for Group Activity Recognition.” Computer Vision and Pattern Recognition. 2016
- [3] Li, Xin, and Mooi Choo Chuah. "SBGAR: Semantics Based Group Activity Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV). 2017.
- [4] Heng Wang, Alexander Kla¨ser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In CVPR, 2011.
- [5] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. CoRR, abs/1412.0767, 2(7):8, 2014.
- [6] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In Advances in neural information processing systems, pages 577–584, 2003.
- [7] Jeffrey Donahue,Lisa Anne Hendricks,Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In ICCV, pages 2625–2634, 2015.
- [8] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. ICCV, 2016.
- [9] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 759–768, 2015.
- [10] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action local- ization. In Proceedings of the IEEE international conference on computer vision, pages 3164–3172, 2015.
- [11] Xiaojiang Peng and Cordelia Schmid. Multi-region two- stream r-cnn for action detection. In European Conference on Computer Vision, pages 744–759. Springer, 2016.
- [12] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolu- tional neural network (t-cnn) for action detection in videos. arXiv preprint arXiv:1703.10664, 2017.