SBGAR: Semantics Based Group Activity Recognition

Xin Li, Mooi Choo Chuah
Department of Computer Science and Engineering
Lehigh University

ICCV17

Introduction

Automatically recognizing human activities in videos is one of the core tasks in the field of computer vision. Compared to the single-person activity recognition task, group activity recognition requires a more robust scheme that can capture correlated individual actions in group activities.

Common existing approaches:

Step 1. Identify individual person in video frames.
Step 2. Track and recognize individual actions.
Step 3. Infer group activities.

Biggest weakness:

High computation time.

Our Contributions:

1. We propose a novel solution, namely **SBGAR**, for group activity recognition.
2. The proposed scheme is **semantics-based**. It can generate a semantic representation for each video frame.
3. Our solution yields **better performance** than state-of-the-art approaches.

Intuition

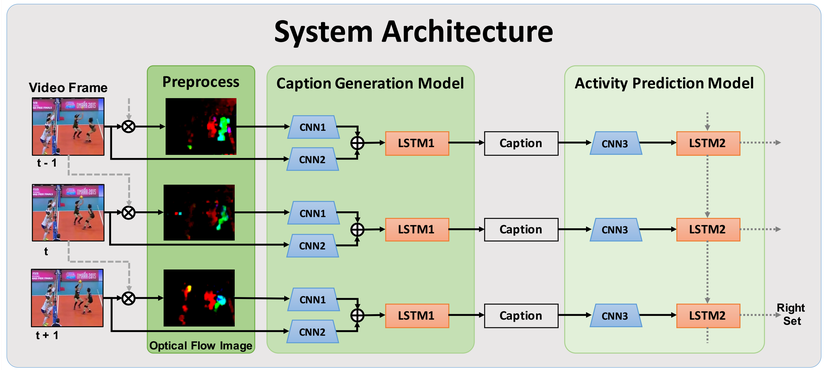
In a Volleyball Game, given the following descriptions:

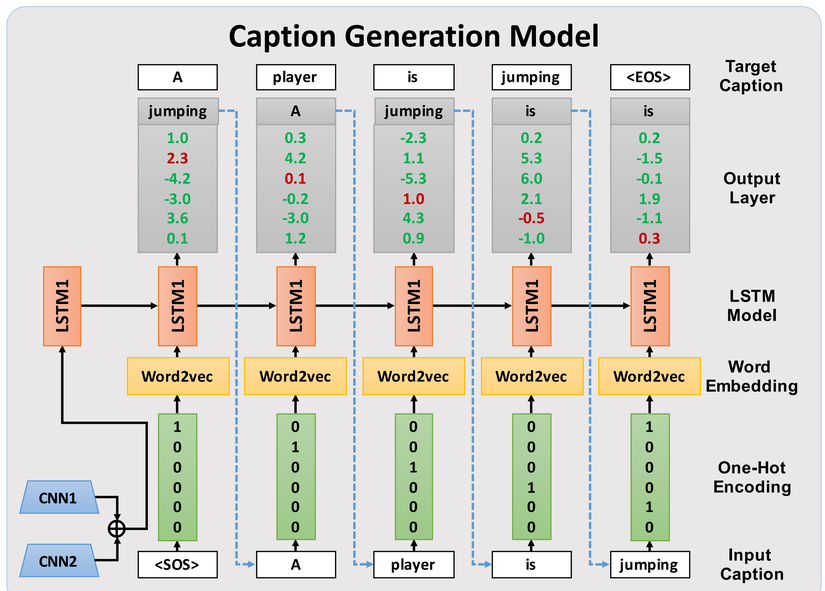
- **Frame t-1:** There is a player **jumping** on the **right** side, while others are **standing**.
- **Frame t:** There is one player **spiking** on the **right** side and three players **blocking** on the **left** side, while others are **standing**.
- **Frame t+1:** All players are **standing**.

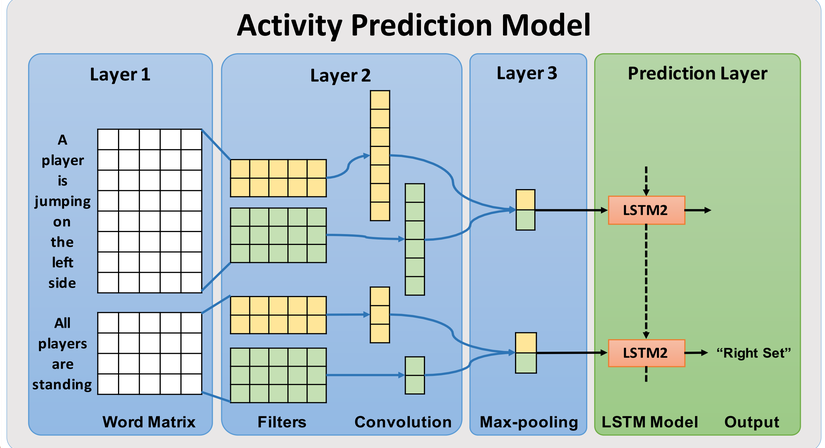
One can easily infer: **Right team is Spiking**.

This work is partially supported by a NSF CSR grant 1217379 and a GPU donated by NVIDIA.

Proposed Solution







Experimental Results

Collective Activity Dataset:

Methods	Accuracy (%)
B1 - Single Frame Classification	67.2
B2 - Temporal Model with Image Features	68.5
B3 - SBGAR (RGB Frame Only)	83.7
B4 - SBGAR (Optical Flow Image Only)	70.1
Contextual Model [1]	79.1
Deep Structured Model [2]	80.6
Two-stage Hierarchical Model [3]	81.5
Cardinality kernel [4]	83.4
SBGAR (RGB & Optical Flow)	**86.1**

Volleyball Dataset:

Methods	Accuracy (%)
B1 - Single Frame Classification	41.9
B2 - Temporal Model with Image Features	44.3
B3 - SBGAR (RGB Frame Only)	38.7
B4 - SBGAR (Optical Flow Image Only)	54.3
Two-stage Hierarchical Model [3]	51.1
SBGAR (RGB & Optical Flow)	**66.9**

Computation Time:

Process	Computation Time (ms)
Optical Flow Image	22.19
Extract CNN1 Feature (Inceptionv3)	27.78
Extract CNN2 Feature (Inceptionv3)	27.78
Caption Generation	28.63
Activity Recognition (10 Frames)	2.15
In Total	108.53

[1] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "discriminativeminative latent models for recognizing contextual group activities," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012.
[2] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidha- ran, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," arXiv preprint arXiv:1506.04191, 2015.
[3] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition." CVPR, 2016.
[4] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi- instance cardinality potential kernel," CVPR, 2015.