LSTA: Long Short-Term Attention for Egocentric Action Recognition

Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz

sudhakaran@fbk.eu sergio@maia.ub.es lanz@fbk.eu

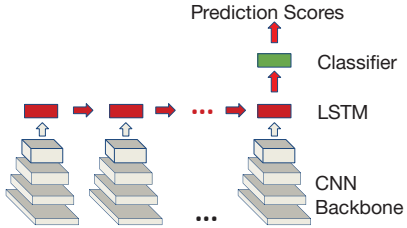
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

1. Goal

Recognition of fine-grained egocentric actions involving object manipulations by an **end-to-end trainable** architecture with **video-level supervision** alone

2. Motivation and Architecture

- Encoding of **long-range temporal** information: **Recurrent Neural Network**
- Encoding of **spatio-temporal** features of relevant regions: **Spatial Attention**

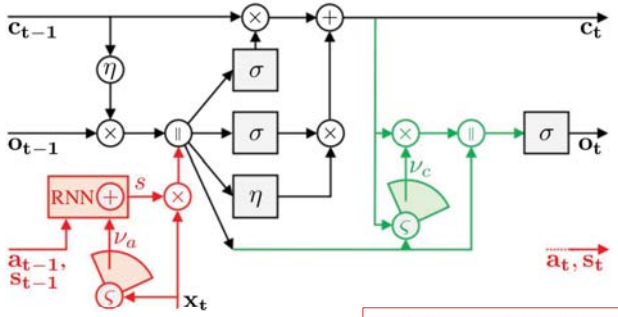


3. Analysis of LSTM

- Fully-connected gates in standard LSTM results in propagation of spatially unstructured memory state: Addressed by **ConvLSTM**
- Spatial features are localized. Attention filtering is performed by gating neurons and requires pre-filtering of input: Addressed by **Spatial Attention**
- Memory tracking is controlled by output gating. Improving the output gating results in better memory propagation: Addressed by **Output Pooling**

LSTA integrates the above solutions into a novel Recurrent Neural Unit

3. Long Short-Term Attention (LSTA)



LSTA enhances LSTM with:

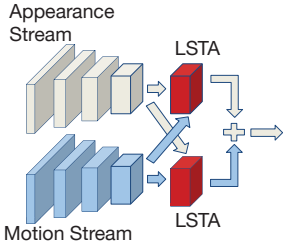
- **In-built recurrent attention**: Weight selector (ς) selects attention weights from a pool of learnable weights. Standard RNN tracks the attention map generated across frames
- **Novel output pooling**: Constraints the model to expose a distilled view of internal memory resulting in a smooth and focused tracking of the latent memory state across the sequence

$$\begin{aligned} \varsigma(\mathbf{x}, \{\theta_c\}) &= \epsilon^{\perp}(\pi(\mathbf{x}, \theta_{c^*})) \\ c^{-} &= \arg\max_{c} \pi(\epsilon(\mathbf{x}), \theta_c) \\ \epsilon(\mathbf{x}) &\leftarrow \text{spatial average pooling} \\ \pi(\epsilon, \theta_c) &\leftarrow \text{linear mapping} \\ \theta &\leftarrow \text{learnable weights} \end{aligned}$$

$$\mathbf{o_t} = \sigma(W_o * [\nu_c \odot \mathbf{c_t}, \mathbf{o_{t-1}} \odot \eta(\mathbf{c_{t-1}})])$$

4. Cross-modal Fusion

- Feature from one stream is used to control the bias of LSTA gates of the other stream
- Motion stream consists of single stack of optical flow



5. Results

Datasets:
1.GTEA61
2.GTEA71
3.EGTEA Gaze+
4.EPIC-Kitchens

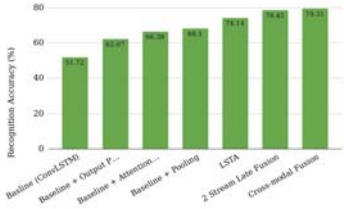


Fig. Ablation Study on the fixed split of GTEA61

Method	GTEA61*	GTEA61	GTEA71	EGTEA	EPIC-S1	EPIC-S2
Li *et al*. [1]	66.8	64	62.1	46.5	-	-
Ma *et al*. [2]	75.08	73.02	73.24	-	-	-
Two Stream	57.64	51.58	49.65	41.84	13.23	7.31
TSN	67.76	69.33	67.23	55.93	20.54	10.89
eleGAtt	59.48	66.77	60.83	57.01	-	-
ego-rnn	77.59	79	77	60.76	-	-
LSTA-RGB	74.14	71.32	66.16	57.94	30.16	15.88
LSTA	**79.31**	**80.01**	**78.14**	**61.86**	**32.60**	**18.71**

Tab. Comparison with state-of-the-art techniques

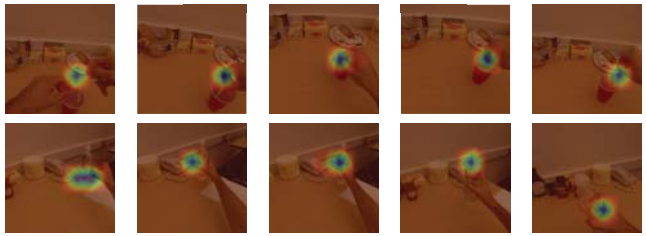


Fig. Attention maps generated by the network

[1] Y. Li, Z. Ye, and J.M Rehg. Delving into Egocentric Actions. CVPR 2015
[2] M. Ma, H. Fan, and K.M. Kitani. Going deeper into first-person activity recognition. CVPR 2016

https://github.com/swathikirans/LSTA