Video Description with Spatial-Temporal Attention

Yunbin Tu[1], Xishan Zhang[2], Bingtao Liu[1] and Chenggang Yan[1]

[1]Hangzhou Dianzi University, [2]Institute of Computing Technology

ACMmultimedia25

Problem

Detail missing

- **Ground truth:** A man is calling.
- **TAT:** A man is talking (on the phone).

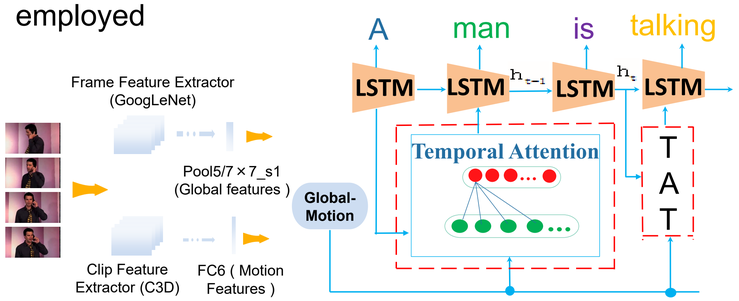
Recognition error

- **Ground truth:** A man is cutting a tree.
- **TAT:** A man is cutting a head.

Motivation

Previous work

➢Only coarse frame-level global-motion features are employed



Our ideal

➢Adding a new feature— object-level local features

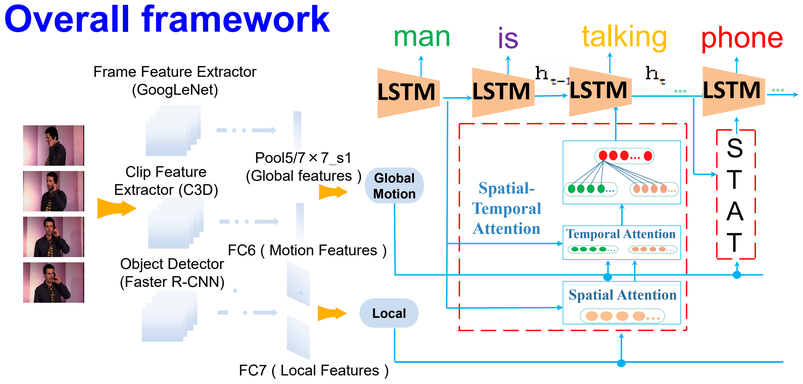
Exploiting local features extracted by Faster R-CNN [19] to address the problem of detail missing.

➢Spatial-temporal attention mechanism

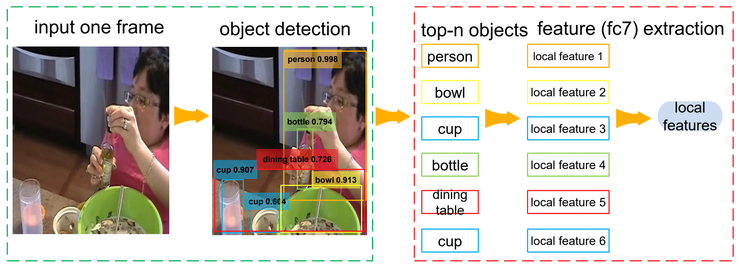
The proposed two-stage attention mechanism can recognize the salient objects more precisely with high recall and automatically focus on the most relevant spatial-temporal segments given the sentence context.

Approach

Overall framework



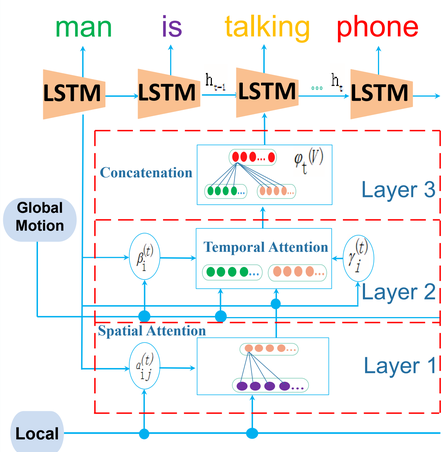
1 Local features extraction



1)Detecting salient objects by pre-trained Faster R-CNN model;

2)Selecting top-n objects as local features by class confidence scores.

2 Spatial-Temporal Attention mechanism



1) Selecting salient objects by spatial attention weights:

$$e_{ij}^{(t)} = w_l^T \tan h(W_e h_{t-1} + U_e v l_{ij} + z_e)$$

$$\alpha_{ij}^{(t)} = \exp\left\{e_{ij}^{(t)}\right\} / \sum_{j\prime=1}^{n} \exp\{e_{ij\prime}^{(t)}\}$$

2) Attending to important frames by temporal attention weights:

$$b_i^{(t)} = w_k^T tanh(W_b h_{t-1} + U_b v[gm]_i + z_b)$$

$$c_i^{(t)} = w_r^T tanh(W_c h_{t-1} + U_c \psi_i(VL) + z_c)$$

$$\gamma_i^{(t)} = \exp\left\{c_i^{(t)}\right\} \Big/ \sum_{i\prime=1}^{k} \exp\{c_{i\prime}^{(t)}\}$$

$$\beta_i^{(t)} = \exp\left\{b_i^{(t)}\right\} / \sum_{i\prime=1}^{k} \exp\{b_{i\prime}^{(t)}\}$$

3) Concatenating global-motion temporal representation and local temporal representation:

$$\varphi_t(V) = \{\varphi_t(VGM), \varphi_t[\psi(VL)]\}$$

Results

MSVD dataset

	B@1	B@2	B@3	B@4	METEOR	CIDEr
TAT-NL (G+C)	0.803	0.676	0.572	0.464	0.318	0.625
NAT (G+C+R-fc7)	0.764	0.627	0.521	0.415	0.315	0.629
TAT (G+C+ R-fc7)	0.773	0.642	0.540	0.432	0.307	0.597
STAT (G+C+ R-fc7)	**0.826**	**0.714**	**0.616**	**0.511**	0.327	**0.675**
TA[37](G+3D CNN)	0.800	0.647	0.526	0.419	0.296	0.517
LSTM-E[15](V+C)	0.788	0.660	0.554	0.453	0.310	-
h-RNN[39](V+C)	0.815	0.704	0.604	0.499	0.326	0.658
HRNE[14](G+C)	0.811	0.686	0.578	0.467	**0.339**	-
M-Fusion[8](V+C)	0.811	0.703	0.607	0.499	0.318	0.634

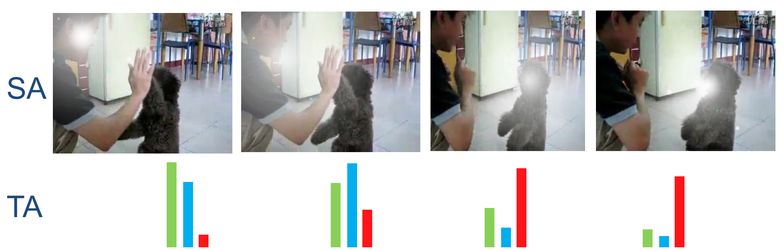
[1] (G=GoogLeNet, C=C3D, R-fc7=Faster R-CNN fc7, V=VGG)

MSR-VTT-10K dataset

	Test split			Valid split		
	B@4	METEOR	CIDEr	B@4	METEOR	CIDEr
TAT-NL (G+C)	0.371	0.264	0.398	0.379	0.269	0.405
NAT (G+C+R-fc7)	0.348	0.250	0.365	0.347	0.252	0.350
TAT (G+C+ R-fc7)	0.343	0.243	0.319	0.358	0.247	0.316
STAT(G+C+ R-fc7)	0.374	0.266	0.415	0.380	0.271	0.402
v2t_nagvigator[6]	**0.408**	0.282	0.448	**0.394**	**0.275**	**0.480**
C3D+Res[18]	-	-	-	0.385	0.267	0.411
SA-LSTM[30]	0.405	**0.299**	-	-	-	-

[1] (G=GoogLeNet, C=C3D, R-fc7=Faster R-CNN fc7)

Qualitative results



Ground truth: A boy is playing with a dog.
TAT: A boy is playing with a baby.
STAT: A boy is playing with a dog.

Ground truth: A woman is applying makeup to her eyes.
TAT: A woman is showing how to make a makeup.
STAT: A woman is applying makeup to her face.