Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description

Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li,and Qi Tian
Institute of Computing Technology,University of Texas at San Antonio

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017

Problem

Description ambiguity:

- **Recognition error**

- **Ground truth:** guys on mat wrestling
- **LSTM:** a group of people are dancing on a track

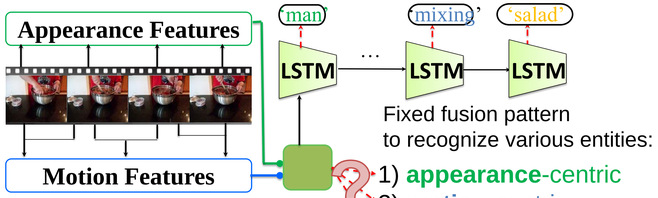
- **Detail deficiency**

- **Ground truth**: a person is mixing a bowl of food
- **LSTM:** a person is cooking

Motivation

Previous work:

- **Weakness inherent in static fusion methods**



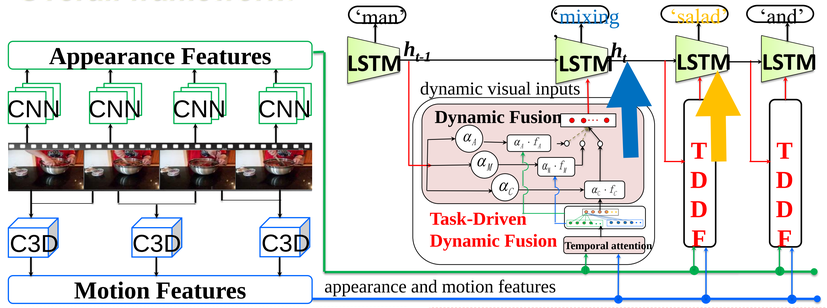
Our idea:

- **Task-driven dynamic fusion**

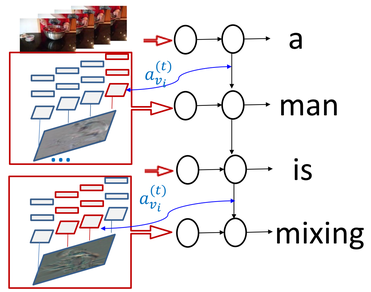
Adaptively choose different fusion patterns according to task status.
Attend to certain visual cues to promote the recognition of
1) **appearance**-centric, 2) **motion**-centric and 3) **correlation**-centric entities.

Approach

Overall framework:



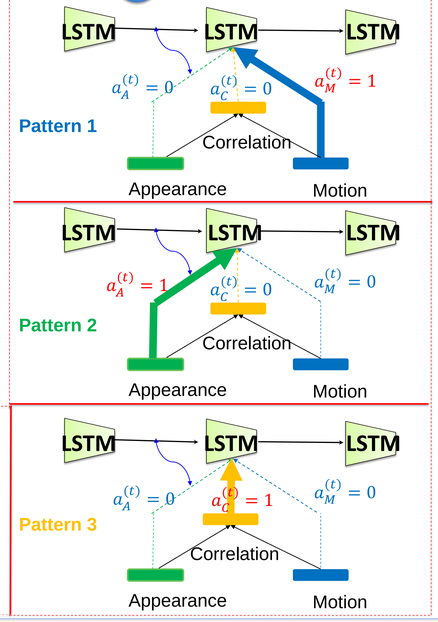
1 Temporal attention



Temporal attention weights:

$$a_{v_i}^{(t)} = f(h_{t-1}, v_i) = W_a \tanh(W_h h_{t-1} + W_v v_i)$$

2 Dynamic fusion



Three different fusion patterns are designed to support the recognition of appearance-centric, motion-centric and correlation-centric entities.

Feature channel fusion weights:

$$a_{p_i}^{(t)} = f_{p_i}(h_{t-1}) = s_{pi}^{(t)} c_{pi}$$

Results

MSVD dataset

	METEOR	(%↑)	CIDEr	(%↑)	BLEU4	(%↑)
VGG	0.302	-	0.563	-	0.416	-
C3D	0.303	-	0.542	-	0.412	-
CON(VGG+C3D)	0.317	(4.6%↑)	0.652	(15.8%↑)	0.428	(2.9%↑)
MAX-2(VGG+C3D)	0.308	(1.7%↑)	0.558	(-0.9%↑)	0.417	(0.2%↑)
SUM-2(VGG+C3D)	0.307	(1.3%↑)	0.654	(16.1%↑)	0.438	(5.3%↑)
MAX-3(VGG+C3D)	0.313	(3.3%↑)	0.663	(17.7%↑)	0.452	(8.6%↑)
SUM-3(VGG+C3D)	0.314	(3.6%↑)	0.602	(6.9%↑)	0.440	(5.8%↑)
TA [34]	0.296	-	0.517	-	0.419	-
LSTM-E [18]	0.310	(3.7%↑)	-	-	0.453	(8.6%↑)
h-RNN [36]	0.326	(4.8%↑)	0.658	(6.0%↑)	**0.499**	(2.2%↑)
HRNE [17]	0.331	-	-	-	0.438	-
TDDF(VGG+C3D)	**0.333**	**(10.0%↑)**	**0.730**	**(29.7%↑)**	0.458	**(10.1%↑)**

MSR-VTT-10K dataset

	BLEU4	(%↑)	METEOR	(%↑)	CIDEr	(%↑)
VGG	0.338	-	0.263	-	0.384	-
C3D	0.363	-	0.263	-	0.397	-
GoogLeNet	0.328	-	0.268	-	0.398	-
v2t_nagvigator [4]	0.408	-	0.282	-	0.448	-
SA-LSTM(VGG+C3D) [10]*	0.405	(0.9%↑)	0.299	(1.7%↑)	-	-
TDDF(GoogLeNet+C3D)	0.372	(2.5%↑)	0.277	(3.3%↑)	0.441	**(10.8%↑)**
TDDF(VGG+C3D)	0.373	**(2.7%↑)**	0.278	**(5.7%↑)**	0.438	(10.3%↑)

Qualitative results

Ground truth: A woman is giving a cooking instructional video
CON: A woman is cooking a dish in a pan
TDDF: A woman adds ingredients to a pot of water

Ground truth: Two football players are running
CON: A person is explaining something
TDDF: Football players playing football

Ground truth: A man is playing with his engine
CON: A man is working on a car
TDDF: A man is fixing a car engine