Spatio-Temporal Ranked-Attention Networks for Video Captioning

Anoop Cherian[1] Jue Wang[2] Chiori Hori[1] Tim K. Marks[1]
[1]Mitsubishi Electric Research Labs, Cambridge, MA [2]Australian National University, Canberra
{cherian,chori,tmarks}@merl.com jue.wang@anu.edu.au

1. Introduction

In this supplementary material, we provide additional qualitative results and visualizations for our Spatio-Temporal and Temporo-Spatial (STaTS) attention based video captioning scheme. We provide several sample captions generated by our model and compare them against ground truth human captions.

2. Qualitative results

In Table 1, we provide examples of video captions generated by our scheme and the human generated captions; for the latter, we randomly selected one caption (out of 20) to show for the respective video. Our provided results are using the STaTS model with I3D features on the MSR-VTT dataset. As can be seen from the results, the captions our scheme generate are very correlated to those by humans (as does our quantitative results show in the main paper).

In Figures 2 and 3, we show qualitative attentions on the respective video frames, the former showing examples when our captions are very similar to human captions, and the latter showing some failure cases. In Figure 4, we show additional results of our ST, TS, and STaTS attention.

Test id#	Reference caption	Generated caption
1 (7517)	a woman is demonstrating various features of a car	a car is being shown
2 (9987)	a finger goes around the corners of a piece of paper	a person is folding a piece of paper
3 (7030)	a ballroom dance class	a group of people are dancing
4 (7519)	optimus prime voice is used briefly during video game play	a man is playing a video game
5 (7518)	a game character is floating in space	a minecraft character is talking
6 (8697)	a boy is sitting on a chair outside he is being recorded while he sings and plays the guitar	a man is singing a song
7 (8696)	a guy swims in blue goggles	a woman is swimming in the water
8 (7886)	a man is demonstrating how to slice a potato thinly using a knife and a cutting board	a man is cutting potatoes
9 (9525)	a chef slices up a fish	a woman is showing how to make a dish
10 (8168)	a guy is playing golf	a man is talking about a dog
11 (8765)	a guy opens a box for a toy car	a man opens a box
12 (9405)	red balloons containing small gifts dropping to the people of the city	a group of people are playing a rocket

Table 1. Captions generated our STaTS model and the corresponding human generated caption for the video. The video id from the MSR-VTT dataset is also shown.

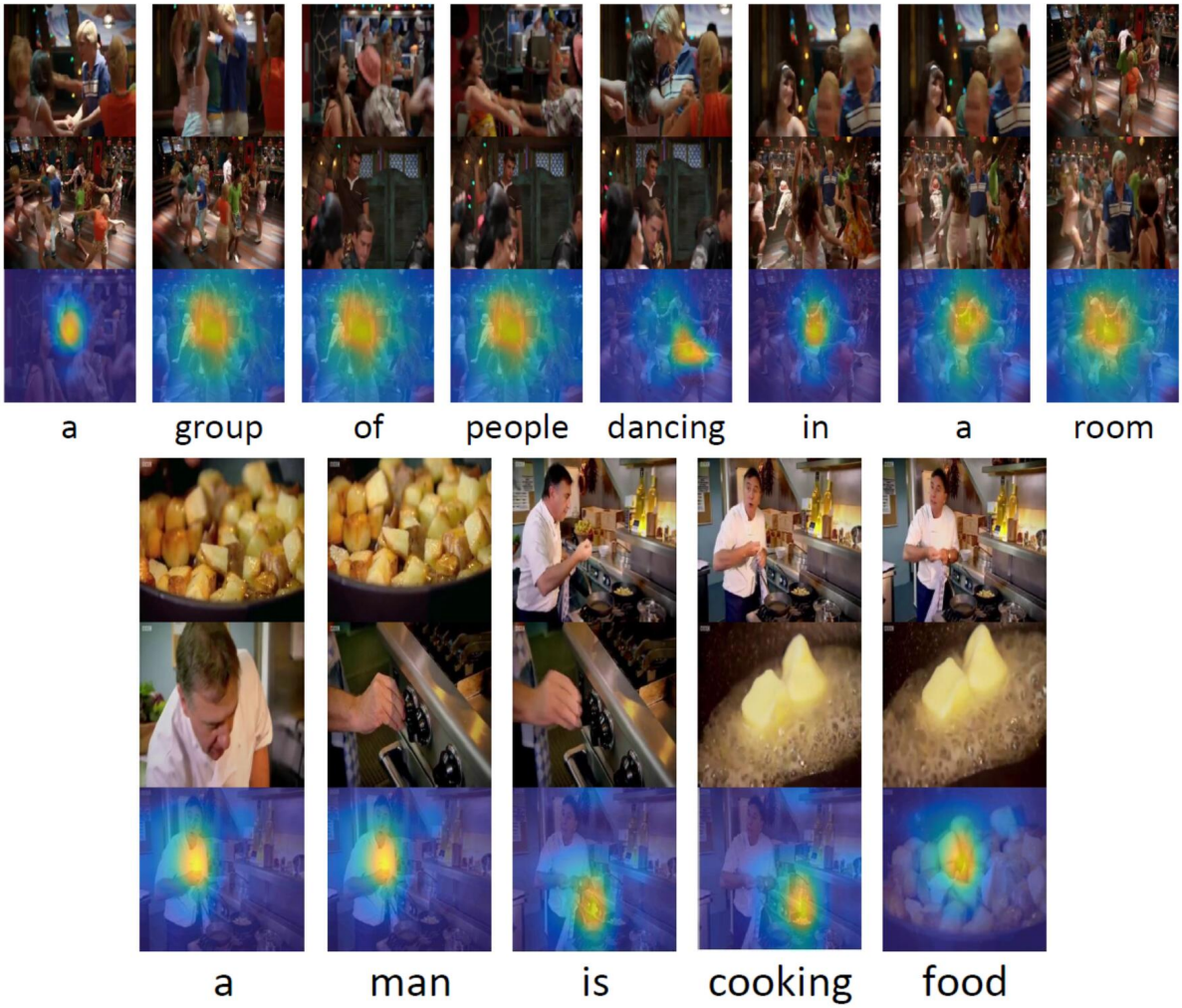


Figure 1. Attention Visualization. The frames in the first two rows (of each sub figure) show the temporal sequence of the video. The 3rd row shows the frames selected by our TS model for each word in the generated caption, overlaid with its corresponding spatial attention map.

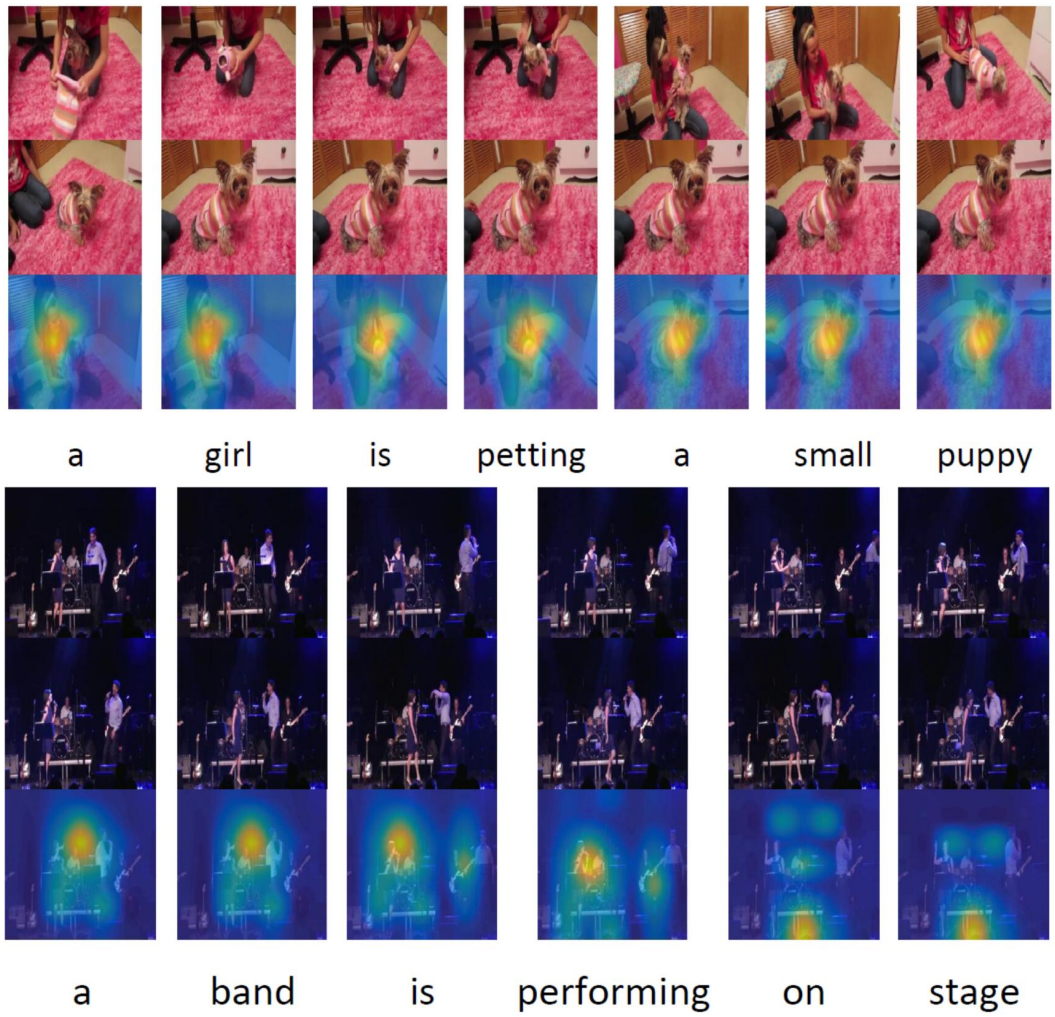
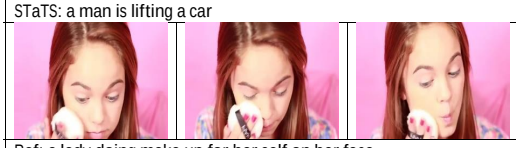


Figure 2. Attention Visualization. The frames in the first two rows show the temporal sequence of the video. The 3rd row shows the frames selected by our TS model for each word in the generated caption, overlaid with its corresponding spatial attention map.

Figure 3. Failure case: the system fails to recognize the sound. The 14 frames in the first two rows show the temporal sequence of the video. The 3rd row shows the frames selected by our TS model for each word in the generated caption, overlaid with its corresponding spatial attention map.

Ref: a man is lifting up a truck	Res: there is a man in white is riding a horse
ST: a man is pushing a car	ST: a man is riding a horse
TS: a man is talking a car	TS: a group of people are running and jumping
STaTS: a man is lifting a car	STaTS: a group of people are riding a horse in a race and a track



Ref: a lady doing make up for her self on her face	Ref: a person showing how food is made
ST: a woman is making up on her face	ST: a man is cooking a dish
TS: a woman is singing	TS: a woman is talking about something
STaTS: a woman is showing how to put makeup on her face	STaTS: a woman is cooking a dish in a kitchen

Ref: a woman is making cuts on a piece of meat with a knife	Ref: the man are in a fight
ST: a woman is cutting a piece of bread	ST: a man is fighting
TS: a person is pouring	TS: two men are dancing
STaTS: a woman is slicing meat	STaTS: two men are fighting

Ref: a man is riding on a horse	Ref: a man bounces a ball once then throws it through a hoop
ST: a woman is riding a horse	ST: a man is playing football
TS: a man is riding a horse	TS: a man is playing a ball
STaTS: a man is riding a horse	STaTS: a man is playing basketball

Figure 4. Qualitative results using our attention model.