Supplementary Materials for End-to-End Video Captioning with Multitask Reinforcement Learning

Anonymous WACV submission

Paper ID 164

In the supplementary materials, we report more qualitative results of our multitask reinforcement end-to-end (E2E) trained model. Table 1 and Table 2 show more qualitative results of different methods on the MSVD dataset [1], and Table 3 and Table 4 are some examples on MSR-VTT [2] dataset.

References

[1] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[2] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.

Table 1. Qualitative results of video captioning on the MSVD dataset (GT: ground truth).

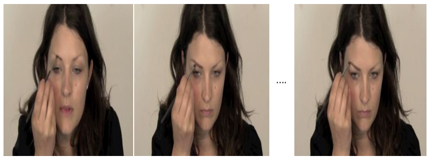
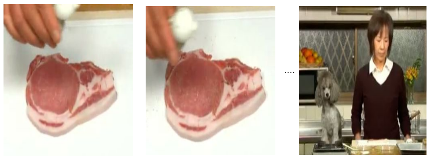
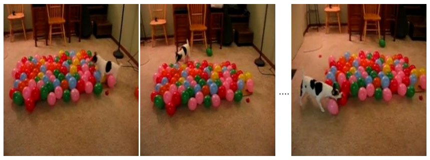
		
S2VT: a woman is talking on a cell	S2VT: a man is putting meat into a pork chop	S2VT: a baby is playing with a ball
E2E: a woman is applying makeup	E2E: a woman is seasoning meat	E2E: a dog is playing with a ball
GT: a woman is applying makeup	GT: someone is seasoning meat	GT: a dog is playing with ball
GT: a woman is putting on her makeup	GT: a person seasons a piece of raw meat	GT: a dog is popping balloons
GT: a woman is painting her eyebrow	GT: a person is seasoning a pork chop	GT: a dog attacks a bunch of balloons

Table 2. Qualitative results of video captioning on the MSVD dataset (GT: ground truth).

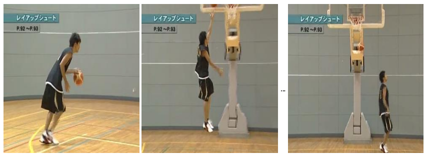
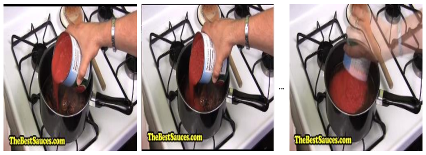
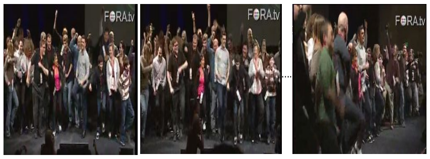
		
S2VT: a boy is playing with a ball	S2VT: a man is pouring a can of water into a pot	S2VT: the men are doing the something
E2E: a man is playing a basketball	E2E: a man is pouring sauce into a pot	E2E: men are dancing
GT: a man is playing basketball	GT: a person pours tomato sauce in a pot	GT: group of people are dancing
GT: a man scores when playing basketball	GT: a man puts tomato sauce in a pot	GT: many members are dancing
GT: a basketball player made a basket	GT: a chef pours sauce into a sausage pan	GT: numerous people are dancing on a stage

Table 3. Qualitative results of video captioning on the MSR-VTT dataset (GT: ground truth).

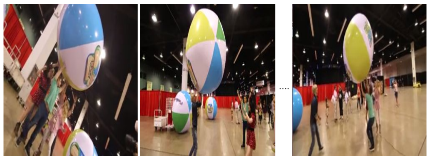
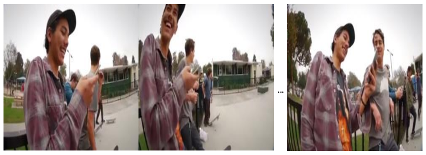
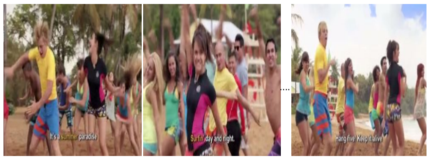
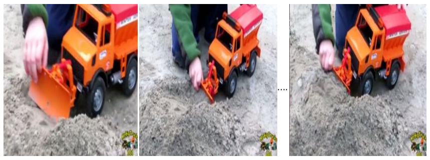
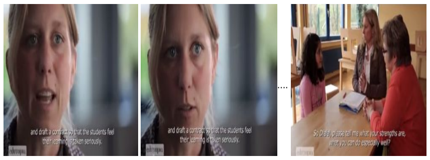
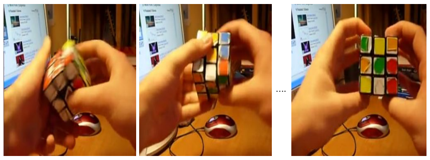
		
S2VT: a man is playing a basketball game	S2VT: a man is singing	S2VT: a group of people are playing in the beach
E2E: a group of people are playing	E2E: a group of people are talking	E2E: a group of people are dancing on the beach
GT: a group of guys playing with a big ball	GT: bunch of guys talking to the camera as they skate board	GT: a group of people dance on the beach
GT: a group of kids is hitting a big beach ball	GT: the man in the plaid shirt and a hat talk to the other man	GT: people are dancing on a beach
GT: people are holding up a beach ball	GT: a group of kids are hanging out at a skate park	GT: teens sing and dance on the beach

Table 4. Qualitative results of video captioning on the MSR-VTT dataset (GT: ground truth).

		
S2VT: a man is playing a car	S2VT: a man is talking about a man	S2VT: a person is showing how to use a device
E2E: a boy is playing a toy	E2E: a woman is talking	E2E: a person is playing with a cube
GT: a boy is playing with a toy truck on a beach	GT: a woman is talking about education	GT: a person is solving a rubix cube
GT: a child plays with a toy truck outside	GT: a woman is talking about students	GT: someone is playing game
GT: a young boy playing with a toy dump truck	GT: a girl discuss to a matter	GT: a person attempts to solve a rubix cube