Supplementary Material for "Towards Diverse Paragraph Captioning for Untrimmed Videos"

Yuqing Song[1]*, Shizhe Chen[2]*, Qin Jin[1]†
[1]Renmin University of China, [2]INRIA
syuqing@ruc.edu.cn, cshizhe@gmail.com, qjin@ruc.edu.cn

1. Human Evaluation

The automatic metrics (METEOR, CIDEr etc.) have been shown to prefer repetitive captions rather than diverse captions in previous works [1]. Therefore, we further conduct human evaluation to show the improvements of our model. Specifically, we randomly select 10% of videos from the ActivityNet ae-test set and hire 10 workers to rank different models from three aspects: accuracy, diversity and coherence. All workers have a good education background and English skills to guarantee that the video content and generated descriptions can be correctly understood. Each video is evaluated by at least two different workers.

Table 1 shows the average rank of our model and other compared baselines, where the MART [2] model generates each sentence with ground-truth event segment annotation. It shows that our model outperforms both the vanilla model and the state-of-the-art MART model on three aspects. The average rank of MART is between 2 and 3, while the vanilla model and our final model rank in the top 2, which demonstrates the advantage of one-stage framework for video paragraph generation. Specially, our model achieves more improvements on the coherence aspect, which benefits from the proposed dynamic video memory. It helps the model to generate the paragraph with a reasonable descriptive logic by controlling the visual focuses on the video sequence.

Table 1. Human evaluation on the ActivityNet ae-test set from three aspects. The lower the rank is, the better the model is.

Method	Average Rank		
	Accuracy	Diversity	Coherence
MART [2]	2.16	2.15	2.43
Vanilla	1.98	1.99	1.93
Ours	**1.85**	**1.86**	**1.63**

[1]Equal contribution. This work was performed when Shizhe Chen was at Renmin University of China.
[2]Corresponding author.

2. High Frequency Phrases List

We present the top 20 frequent verb phrases in the ActivityNet annotations of training set in Table 2. Half of them are related to the "camera" or "screen", which can be used to describe multiple videos and convey limited unique information of different events. However, they are frequently generated by the language decoder due to the priors in the data distribution. Besides their frequencies in the ground-truth, we also show their frequencies in the generated paragraphs from vanilla baseline model and our final model in the table. With our proposed dynamic video memories and diversity-driven training strategies, our model generates less uninformative high-frequency phrases (colored in red), and more video-related phrases (colored in blue).

Table 2. The top 20 frequent verb phrases in ActivityNet dataset and their occurring frequencies (%) in the ground-truth and generated paragraphs from vanilla model and our final model.

#	Verb Phrase	GT	Vanilla	Ours
1	speaking/talking to the camera	15.77	42.82	40.74
2	watch on the side	3.60	6.80	2.73
3	stands/gets up	3.06	2.08	0.97
4	appears on the screen	2.67	0.00	0.08
5	looking to the camera	1.85	11.27	0.16
6	smiling to the camera	1.71	3.58	0.73
7	see the ending screen	1.29	0.00	0.00
8	walking into frame	1.24	1.87	1.50
9	speaking to one another	1.19	1.42	0.94
10	camera pans (all) around	1.00	9.77	3.83
11	see an opening title screen	0.97	0.00	0.00
12	falls to the ground	0.89	0.04	0.08
13	walks away	0.82	3.95	4.76
14	hit the ball	0.77	1.14	2.65
15	walk out of frame	0.76	0.20	0.04
16	playing the drums	0.67	0.98	1.55
17	running down the track	0.66	0.61	0.53
18	sitting down	0.61	0.08	1.06
19	bends down	0.59	0.49	1.30
20	looking off into the distance	0.58	2.97	0.16

VTransformer (GT events)
A man is chopping wood in a forest. A man is chopping wood in a forest.

AdvInf (GT events)
A man is seen holding a ax and holding a stick in his hands. He then swings the axe and misses the stick.

MART (GT events)
A man is chopping wood in a field. He swings the ax at a piece of wood.

Vanilla (no event detection)
A man is seen using a piece of wood and pushing a piece of wood with a stick. He swings the stick around and throws the stick up in the air.

Ours (no event detection)
A man is seen standing before a log and holding an ax. The man then uses the ax to split the log in half while the camera pans around.

Ground-Truth
A man is seen standing in the woods holding an ax and walking closer to a log. The man steps on the log a bit followed by him swinging the ax on the log.

VTransformer (GT events)
The stick falls down and the ground in the ground. She hits the ball with the bat. The woman hits the ball to her ball.

AdvInf (GT events)
A man is standing on a field behind a net. The woman in black shirt is standing next to the ball. The man in the red shirt hits the ball.

MART (GT events)
A man in a red shirt is playing a game of cricket. He hits the ball with it and hits it. He hits the ball and it in slow motion.

Vanilla (no event detection)
A man is seen playing a set of bagpipes while people walk around him and watch. The man continues to play as the people walk around and out of frame as well as him.

Ours (no event detection)
A game of croquet is shown with a man holding a stick and hitting the ball with the stick. The man hits the ball very hard and then it ends up hitting it off into the distance.

Ground-Truth
A boy walks backwards on the lawn. The boy picks up the croquet stick and makes his shot. The boy smiles and heads towards the ball.

Figure 1. Qualitative examples of the generated paragraphs from our model and other state-of-the-arts methods. The red represents wrong descriptions, the yellow represents repetitive descriptions and the blue represents the correct description generated by our model.

3. Additional Qualitative Examples

We visualize some examples of the generated paragraphs from our model and state-of-the-arts methods in Figure 1 and Figure 2. Figure 1 shows that our model can generate more accurate and diverse paragraphs compared with both the vanilla baseline and other state-of-the-art models which even use ground-truth event boundary annotations. We also show a failure example in Figure 2. In this example, though our model can generate diverse and coherent paragraph, the last sentence is not relevant to the video content. Furthermore, the models with ground-truth temporal event annotations can correctly describe the background clips in the video, while our model usually ignores them. With more annotated events, they also generate longer paragraphs while our model tends to describe the video with 2 or 3 sentences.

References

[1] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: story oriented dense video captioning evaluation framework. In *European Conference on Computer Vision*, pages 517–531, 2020. 1

[2] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. MART: memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614, 2020. 1

VTransformer (GT events)
A group of men are in a gym. Two people are boxing in a boxing room. There is a person dressed in a boxing gloves. There is a woman dressed in a boxing gloves. They then show that they begin to fight. The boxer in the black and red shorts punches the boxer in the black shorts. There are several spectators watching them perform. The coach is standing next to them.

AdvInf (GT events)
A black screen appears with white text that says www art dot com instant for a video about how to do kickboxing. The two men are in boxing gloves and punches with one another. The boxer in red and black shorts kick the ball with his partner that has been selected for his opponent to make a big punch kick. The man in black kicks and punches the punches and punches the punches. The two continue to fight with one another and punches for him. The men shake hands and hug. The two men kick each other on the mat. The video ends with the closing credits shown on the screen.

MART (GT events)
The credits of the clip are shown. Two men are boxing in a boxing ring. The men kick and punch at each other. The men are boxing and kicking. The men are boxing with one another. The men are boxing with one another. The men are boxing and kicking. The red team is seen in the air.

Vanilla (no event detection)
A man is seen speaking to the camera while a large group of people watch him on the side. The man then begin kicking the man back and fourth. The man continues to fight while moving his arms around and legs.

Ours (no event detection)
Two men are boxing in a boxing ring. They are kicking and punching at each other. They finish and hug each other.

Ground-Truth
We see the black title screen. We then see two people boxing in a gym. The screen goes wavy briefly. The screen changes and the lady is wearing protective gear as the man practices. We see the man boxing with another man. The men pause and begin again. The screen goes black and we see kids hitting punching bags. We see an image of the boy, and the closing screen.

Figure 2. A failure example of our proposed model, which ignores background clips in the video and generates shorter paragraphs. The words in blue represent background clip descriptions.