End-to-End Dense Video Captioning with Parallel Decoding (Supplementary Materials)

Teng Wang^{1,2}, Ruimao Zhang^{3,4}, Zhichao Lu², Feng Zheng^{2*}, Ran Cheng², Ping Luo¹ ¹ The University of Hong Kong ² Southern University of Science and Technology ³ The Chinese University of Hong Kong (Shenzhen) ⁴ Shenzhen Research Institute of Big Data

tengwang@connect.hku.hk ruimao.zhang@ieee.org luzhichaocn@gmail.com
 f.zheng@ieee.org ranchengcn@gmail.com pluo@cs.hku.hk

A. More Implementation Details

Event proposal generation module based on merely captioning supervision. In Sec. 4.3, we make the following modifications to train an event proposal generation module without localization supervision: 1) We extend the 1D reference point to the 2D reference point $p_j = (p_j^{c}, p_j^{l})$, where p_i^c, p_i^l denote the center and the length of the reference point, respectively. 2) For each decoder layer, we fix the sampling keys in deformable attention as K = 4 evenly spaced positions over a specified interval from $p_i^c - 0.5 p_i^l$ to $p_i^c + 0.5p_i^l$ to stabilize the network training. 3) Without gIOU cost in bipartite matching, it is hard to accurately assign the target captions to event queries. We design the caption cost to mitigate this problem. Given any groundtruth caption $S_{j'} = \{w_{j't}\}_{t=1}^{M_{j'}}$ and any event query features \tilde{q}_j , we obtain the output probabilities $\{c_{jj't}^{cap}\}_{t=1}^{M_{j'}}$ predicted by the captioning head with teacher forcing, where $M_{j'}$ denotes the caption length. The caption cost matrix is calculated by:

$$(C_{\text{cap}})_{jj'} = \frac{1}{M_{j'}^{\gamma}} \sum_{t=1}^{M_{j'}} \log(c_{jj't}^{\text{cap}}),$$

where $\gamma = 2$ is the modulation factor of the caption length. The final cost matrix for bipartite matching is:

$$C = C_{\rm cap} + \alpha_{\rm cls} L_{\rm cls},$$

where $\alpha_{cls} = 0.5$ is the balance factor.

Based on the above modification, we train PDVC_light with merely captioning loss on YouCook2. We choose the lightweight captioning head to ease the optimization difficulty. During inference, we directly use the reference points in the last layer as the predicted proposals.



(a) Predicted Proposals on ActivityNet (b) GT Proposals on ActivityNet Cap-Captions tions



(c) Predicted Proposals on YouCook2 (d) GT Proposals on YouCook2

Figure A1. The distribution of predicted proposals and groundtruth proposals. Horizontal and vertical axes represent the normalized center position and normalized length of proposals, respectively. For each dataset, we report the results of 200 randomly sampled videos on the validation set. The sub-figure (a)/(c) contain 10/100 clusters with different colors, where each cluster corresponds to one event query.

B. Visualization

Predicted proposals. We visualize the distribution of generated proposals of PDVC in Fig. A1. For the ActivityNet Captions dataset, ground-truth proposals are distributed evenly across different positions and different lengths. However, for YouCook2, the length of most ground-truth proposals is relatively small (less than 25% of the video duration). From the figure, we conclude that: 1) Each query describes a specific mode of the proposals' location. 2) All



Figure A2. Dense captioning performance of PDVC on different activity classes. Activity labels are from the ActivityNet1.3 dataset [2].

e1 e2	
	e3
61 62 0	ез
P2	e1
 Ground Truth e1: A man is standing in a room. e2: He has a ball on a tennis racket. e3: He throws the ball in the air and hits it with the racket. MT e1: a man is seen standing on a court holding a tennis racket. e2: a man is standing on a court. e3: the man serves the ball with the racket. 	PDVC_light e1: he throws the ball back and forth. e2: he is then seen spinning around and throwing a ball. e3: he throws the ball back and forth. PDVC e1: a man is standing on a court. e2: a man is seen standing on a tennis court holding a tennis racket. e3: the man then serves the ball and hits the ball.
e1 •	e2 e3
eı ei	e2 e3
e1	e2 e3
Ground Truth e1: A close up of a candle is shown as well as a picture of a man praying in the desert. e2: A person is then seen taking off a pair of shoes in front of him. e3: The man sets the shots in between his legs and speaks to the camera.	PDVC_light e1: a person is seen kneeling down on a table and begins to the camera. e2: the person is then seen <u>putting a tire on the floor</u> and begins to the camera. e3: the person is then shown on the floor.
MT e1: a close up of a <unk> is shown followed by a person walking into frame e2: the man then begins to <unk> the dog 's leg. e3: the man then grabs a pair of <unk> and begins to <unk> the shoe</unk></unk></unk></unk>	PDVC e1: a close up of a piece of shoes are shown followed by a person putting a piece of shoes. e2: the person is then seen putting the shoes on the ground

e3: the man then grabs a pair of <unk> and begins to <unk> the shoe. e3: the man then puts the shoes on the ground. e3: the man then puts the shoes on the shoes and begins to the camera.

Figure A3. Visualization of predicted dense captions. Incorrect phases are underlined in red and the correct ones in green.

queries can predict video-wide proposals with coherence and low redundancy and generate a similar distribution with ground truth. 3) Event queries serve as a location prior for localization tasks, which are trained to learn location patterns of events from human annotations.

Activity types. The dense captioning performance of PDVC varies in different activity types. Fig. A2 shows the METEOR score of PDVC with predicted/ground-truth proposals on 200 activity classes. Our model seems to generate more accurate captions with activities containing distinct scene cues or large objects, like "riding bumper cars", "playing squash", and "calf roping". However, activities that rely more on fine-grained action cues or small objects tend to get a worse METEOR, like "doing karate", "gargling mouthwash", and "rollerblading". It is promising to achieve a performance improvement to incorporate the fine-grained object features and a more powerful action recognition model.

Temporally-localized captions. Fig. A3 shows the generated captions with their temporal locations of different models. The captions of MT [1] are generated based on ground-truth proposals, while PDVC_light and PDVC are

with predicted proposals. For the second video, MT and PDVC_light misrecognize the shoes as a dog and a tire, respectively. Instead, PDVC can generate accurate and meaningful captions with predicted proposals, which verifies the effectiveness of the proposed parallel decoding mechanism and the captioning head with deformable soft attention.

References

- L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "Endto-end dense video captioning with masked transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8739-8748.
- [2] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961-970. 2