Supplementary Materials of SCANet: Scene Complexity Aware Network for Weakly-Supervised Video Moment Retrieval

1. Limitations

Our proposed method is based on the scene complexity of video by referring the number of annotated queries to the video. However, in the real environment, it may not be available to get scene complexity of video from referring to other annotated queries (*i.e*. In real environment, we may not access to other annotated query sets for one video). We feel this is our current SCANet's limitation, and to overcome this, we further made another effort to learn scene complexity via the neural network from the input of video, where we refer to this method as 'Scene Complexity Neural Estimator'. In Section 2, we elaborate this with our current studies as another our experimental contribution.

2. Scene Complexity Neural Estimator

In order for the scene complexity not to depend on the queries attached to the video in an inference time, we introduce a neural network to predict scene complexity of video. We define this network as Scene Complexity Neural Estimator (SCNE). In training, SCNE takes input of video V and produce the discrete probability distribution $p = \{p[1], p[2], \cdots, p[K]\}$, where K is maximum discrete number of scene complexity and $p[i]$ denotes the probability when the scene complexity equals to i. (*i.e*. $\alpha = i$) like:

$$p = f_{sc}^{\theta}(V) \in \mathbb{R}^K, \tag{1}$$

where the f_{sc}^{θ} is SCNE and θ is the learnable parameter. To train the SCNE, we use the the scene complexity of $\alpha^c = f_{sc}(V, D)$ in Equation (1)[1] of the main paper as ground-truth scene complexity. The training loss is the cross entropy loss like below:

$$\mathcal{L}_{scne} = -\log(p[\alpha^c]), \tag{2}$$

where α^c is the ground-truth scene complexity. We use argmax value from the predictions of SCNE to give scene complexity of our SCANet. Table 1 summarizes the retrieval performances of SCANet with SCNE on Charades-STA and ActivityNet Caption dataset. In the method of predicting scene complexity with a neural network, there was a slight decrease in performance due to inaccuracy of prediction (*i.e*. current SCNE accuracy is about 66%.), but SCNE has the advantage of not depending on the queries annotated to the video anymore. Our further research includes the studies to improve the accuracy of scene complexity estimation of SCNE.

[1]Here, we add superscript c as α^c to denote the ground-truth.

Table 1: Performances of applying Scene Complexity Neural Estimator (scne) of SCANet on the Charades-STA (test) and ActivtyNet Caption (val_2) dataset.

Charades-STA						
Method	R@1,IoU=m			R@5,IoU=m		
	m=0.3	m=0.5	m=0.7	m=0.3	m=0.5	m=0.7
SCANet	68.04	50.85	24.07	98.24	86.32	53.28
SCANet (scne)	66.82	50.21	23.41	97.89	84.97	52.86
ActivityNet Caption						
Method	R@1,IoU=m			R@5,IoU=m		
	m=0.1	m=0.3	m=0.5	m=0.1	m=0.3	m=0.5
SCANet	83.62	56.07	31.52	94.36	82.34	64.06
SCANet (scne)	82.51	55.21	31.14	93.81	81.72	63.62

3. Additional results

We provide further experimental results of our proposed SCANet in terms of proposed adaptive proposals and some failure cases.

Candidate Proposals. In Figure 1, we further illustrate the other proposals generated in the same sample of Section 4.5. We configure the proposals by rank from top to bottom. Figure 1(a) is the case when the video includes many scenes. For the given query as 'Person awakens in their home office.', our proposed Complexity-Adaptive Proposal Generation (CPG) determines the number of proposals as $p_\alpha = 9$ by referring to the high value of scene complexity $\alpha = 8$. Among the proposals, it can be seen that proposals with high rank hold a high overlap with the ground-truth. Here, we make proposals with high overlap with a darker color. Furthermore, the adaptive proposals are also capturing the other scenes in the video as other candidates, where we use different colors for them to give distinctions. For

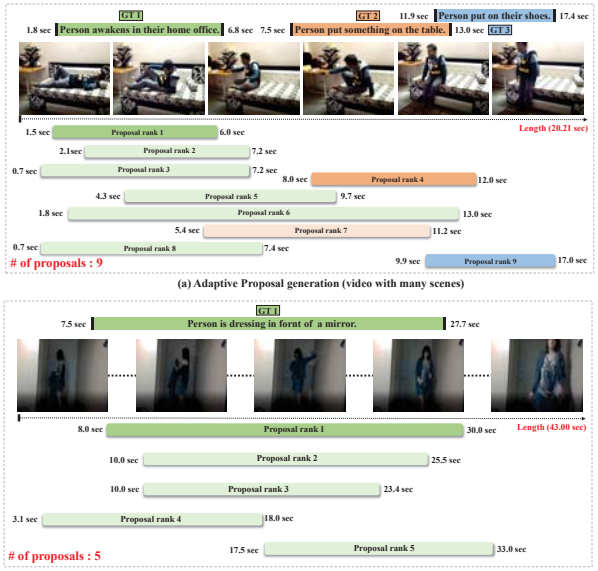

Figure 1: Illustartion of generated proposals from (a) video with many scenes and (b) video with few scene.

the case of Figure 1(b), there is a single scene, where the SCANet provides only 5 proposals by the CPG. This is because the CPG considers the small value of scene complexity $\alpha = 2$. These proposals also overlap with the ground-truth. With these results, it can be confirmed that the proposals are adaptively generated according to the scene complexity and no longer depends on the video length.

References