Supplementary Material for CVPR 2018 paper #330

Chih-Yao Ma*[1], Asim Kadav[2], Iain Melvin[2], Zsolt Kira[3], Ghassan AlRegib[1], and Hans Peter Graf[2]

[1]Georgia Institute of Technology, [2]NEC Laboratories America, [3]Georgia Tech Research Institute

1. Supplementary

1.1. Qualitative analysis on Kinetics

To further validate the proposed method, we qualitatively show how the SINet selectively attends to various regions with relationships and interactions across time. We show several examples in Figure 3, 4, and 5. In each of the figure, the top row of each video frame has generally multiple ROIs with three colors: red, green, and blue. ROIs with the same color indicates that there exist inter-relationships. We then model the interaction between groups of ROIs across different colors. The color of each bounding box is weighted by the attention generated by the proposed method. Thus, if some ROIs are not important, they will have smaller weights and will not be shown on the image. The same weights are then used to set the transparent ratio for each ROI. The brighter the region is, the more important the ROI is.

Focus on object semantics Recent state-of-the-art methods for action recognition rely on single compact representation of the scene. We show that the proposed SINet can focus on the details of the scene and neglect the visual content that maybe irrelevant such as the background information. For example, in Figure 3, the model constantly focus on the rope above the water and the person riding on wakeboard. The same goes for Figure 4. The background scenes with ice and snow are ignored throughout the video since it's ambiguous and easy to be confused with other classes involve snow in the scene.

Adjustable inter-relationships selection We notice that our SINet tends to explore the whole scene early in the video, i.e. the attentions tend to be distributed to the ROIs that cover large portion of the video frame, and the attentions become more focused after this exploration stage.

1.2. Qualitative analysis on ActivityNet Captions

In addition to the qualitative analysis on action recognition task, we now present the analysis on video captioning. Several examples are shown in Figure 6, 7, and 8. At each word generation step, the SINet-Caption uses the weighted sum of the video frame representations and the weighted sum of object interactions at corresponding timesteps (co-attention). Note that, since we aggregate the detected object interactions via the LSTM cell through time, the feature representation of the object interactions at each timestep can be seen as a fusion of interactions at the present and past time. Thus, if temporal attention has highest weight on $t = 3$, it may actually attend to the interaction aggregated from $t = 1$ to $t = 3$. Nonetheless, we only show the video frame with highest temporal attention for convenience. We use red and blue to represent the two selected sets of objects ($K = 2$).

In each of the figures, the video frames (with maximum temporal attention) at different timesteps are shown along with each word generation. All ROIs in the top or bottom images are weighted with their attention weights. In the top image, ROIs with weighted bounding box edges are shown, whereas, in the bottom image, we set the transparent ratio equal to the weight of each ROI. The brighter the region is, the more important the ROI is. Therefore, less important ROIs (with smaller attention weights) will disappear in the top image and be completely black in the bottom image. When generating a word, we traverse the selection of beam search at each timestep.

As shown in Figure 6, we can see that the SINet-Caption can successfully identify the person and the wakeboard. These selections of the two most important objects imply that the person is riding on the wakeboard — water skiing. We also observe that, in Figure 7, the proposed method focuses on the bounding boxes containing both person and the camel. Suggesting that this is a video for people sitting on a camel. However, it failed to identify that there are in fact multiple people in the scene and there are two camels. On the other hand, the SINet-Caption is able to identify the fact that there are two persons playing racquetball in Figure 8.

*work performed as a NEC Labs intern

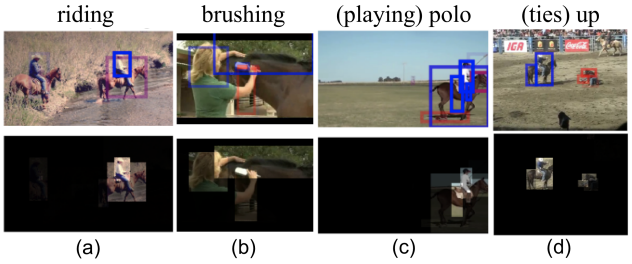

Figure 1. What interactions (verb) learned for video captioning. We verify how the SINet-Caption distinguishes various type of interactions with a common object - *horse*. (a) People are *riding* horses. (b) A woman is *brushing* a horse. (c) People are playing *polo* on a field. (d) The man *ties* up the calf.

1.2.1 Distinguish interactions when common objects presented

A common problem with the state-of-the-art captioning models is that they often lack the understanding of the relationships and interactions between objects, and this is oftentimes the result of dataset bias. For instance, when the model detects both person and a horse. The caption predictions are very likely to be: A man is riding on a horse, regardless whether if this person has different types of interactions with the horse.

We are thus interested in finding out whether if the proposed method has the ability to distinguish different types of interactions when common objects are presented in the scene. In Figure 1, each video shares a common object in the scene - *horse*. We show the verb (interaction) extracted from a complete sentence as captured by our proposed method.

- People are *riding* horses.
- A woman is *brushing* a horse.
- People are playing *polo* on a field.
- The man *ties* up the calf.

While all videos involve horses in the scene, our method successfully distinguishes the interactions of the human and the horse.

1.2.2 Discussion on ActivityNet Captions

We observed that while higher-order object interactions did contribute to higher performance on ActivityNet, the contributions were not as significant as when applied to the Kinetics dataset (quantitatively or qualitatively). We hereby discuss some potential reasons and challenges on applying SINet-Caption on the ActivityNet Captions dataset.

Word by word caption generation: In line with the work from question-answering, machine translation, and captioning, we generate a language sentence describing a video one word after another. At each word generation step, the SINet-Caption uses the last generated word, video frame representations, and their corresponding object interactions. As we can see from both qualitative results from Kinetics and ActivityNet Captions, our proposed method is able to identify the interactions within a very few video frames. However, taking Figure 7 as an example, at the first word "a", our model has already successfully selected the persons (both in light blue and red) on top of the camel (bright red). Yet, during the following caption generation, the SINet-Caption was *forced* to look at the visual content again and again. Introducing the gated mechanism [3] may mitigate this issue, but our preliminary results do not show improvement. Further experiments toward this direction may be needed.

Semantically different captions exist: Each video in the ActivityNet Captions dataset consists of 3.65 (average) different temporal video segments and their own ground truth captions [2]. These video captions have different semantic meanings but oftentimes share very similar video content, i.e. the same/similar video content has several different ground truth annotations. As a result, it may create confusion during the training of the model. Again, taking Figure 7 as an example, we observed that the SINet-Caption often focuses on the person who leads the camels ($t = 1, 3, 15$). We conjecture that this is due to the fact that, within the same video, there exists another video segment with annotation: *A short person that is leading the camels turns around.* Although within the same video content, one of the ground truth focuses on the persons sitting on the camels, another ground truth focuses on the person leading the camels. This seems to be the reason why the trained network focuses on that particular person. Based on this observation, we believe that future work in re-formulating these semantically different annotations of similar video content for network training is needed, and perhaps it may be a better way to fully take advantage of fine-grained object interactions detected from SINet-Caption. One possibility will be associating semantically different video captions with different region-sequences within a video [4].

1.3. Performance improvement analysis on Kinetics

The proposed SINet ($K = 3$) shows more than 5% improvement on top-1 accuracy in 136/400 classes and more than 10% improvement in 46 classes over baseline. We show the classes that were improved more than 10% on top-1 accuracy in Figure 2. In addition to these classes, the proposed SINet in modeling fine-grained interactions specifically improved many closely related classes.

- 7 classes related to **hair** that are ambiguous among each other: *braiding hair*, *brushing hair*, *curling hair*, *dying hair*, *fixing hair*, *getting a haircut*, and *washing*

hair. We show 21% top-1 improvement on *washing hair*; 16% improvement on *getting a haircut*.

- 4 classes related to **basketball** require the model to identify how the basketball are being interacted. These classes are: *playing basketball*, *dribbling basketball*, *dunking basketball*, and *shooting basketball*. We observed 18%, 10%, 6%, and 8% improvement respectively.

- Among 3 related to **juggling** actions: *juggling fire*, *juggling balls*, and *contact juggling*. We obtained 16%, 14%, and 13% improvement respectively.

- Our model significantly improved the **eating** classes, which are considered to be the hardest [1], because they require distinguishing what is being eaten (interacted). We show improvement among all eating classes, including *eating hot dog*, *eating chips*, *eating doughnuts*, *eating carrots*, *eating watermelon*, and *eating cake*. We obtained 16%, 16%, 14%, 8%, 4%, and 4% improvement respectively.

1.4. ActivityNet Captions on 1st and 2nd val set

We report the performance of SINet-Caption on the 1st and the 2nd validation set in Table 1. We can see that using fine-grained (higher-order) object interactions for caption generation consistently shows better performance than using coarse-grained image representation, though the difference is relatively minor compared to the results on Kinetics. We discuss the potential reasons in Sec. 1.2. Combining both coarse- and fine-grained improve the performance across all evaluation metrics. Interestingly, using co-attention on detected object interactions shows better performance on the 1st validation set but has similar performance on the 2nd validation set.

1.5. Model architecture and FLOP

We now describe the model architecture of the proposed recurrent higher-order module and how the FLOP is calculated.

SINet architecture: We first project the image representations $v_{c,t}$ to introduce learnable feature representations. The MLP g_ϕ consist of two sets of fully-connected layers each with batch normalization and ReLU. It maintains same dimension ($m = 2048$) of the input image feature. Thus, the coarse-grained representation of the video is a feature vector with 2048 dimension. Inside the Recurrent HOI module, each of the MLP g_{θ_k} has three sets of batch normalization layers, fully-connected layers, and ReLUs. In the experiments with two attentive selection module ($K = 2$), we set the dimension of the fully-connected layer to be 2048. The concatenation of $v^1_{o,t}$ and $v^2_{o,t}$ is then used as the input to the following LSTM cell. Empirically, we find out that it's important to maintain high dimensionality for the input to LSTM cell. We adjust the dimension of hidden layers in g_{θ_k} given the number of K, e.g. we reduce the dimension of the hidden layer if K increases. In this way, the inputs to LSTM cell have the same or similar feature dimension for fair experimental comparison. The hidden dimension of the LSTM cell is set to be 2048. Before concatenating the coarse- (v_c) and fine-grained ($v_{oi,T}$) video representations, we re-normalize the feature vector with batch normalization layer separately. The final classifier then projects the concatenated feature representation to 400 action classes.

SINet-Caption architecture: We first use a single fully-connected layer with batch normalization, dropout, and ReLU to project the pre-saved image features $v_{c,t}$. The g_ϕ maps the feature vector from 2048 to 1024. We use two attentive selection modules for video captioning task ($K = 2$). Each g_{θ_k} consist of a batch normalization, fully-connected layer, dropout layer, and a ReLU. It maps input object feature vector from 2048 to 512. The dropout ratio for both g_ϕ and g_{θ_k} are set to be 0.5. The concatenation of $v^1_{o,t}$ and $v^2_{o,t}$ is used as input to the LSTM cell inside Recurrent HOI module. The hidden dimension of this LSTM cell is set to be 1024. The dimension of word embedding is 512. We use ReLU and dropout layer after embedding layer with dropout ratio 0.25. The hidden dimension of both Attention LSTM and Language LSTM are set to be 512.

FLOP is computed per video and the maximum number of objects per frame is set to 15. We compare the computed FLOP with traditional object interactions by paring all possible objects. The results are shown in Table 2.

References

[1] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 5

[3] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[4] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue. Weakly supervised dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

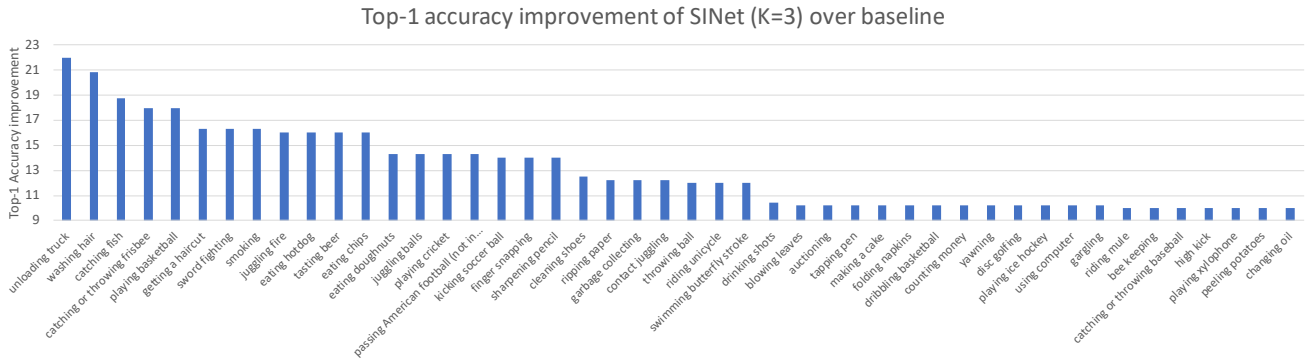

Figure 2. Top-1 accuracy improvement of SINet ($K = 3$) over baseline. 46/400 classes that are improved more than 10% are shown.

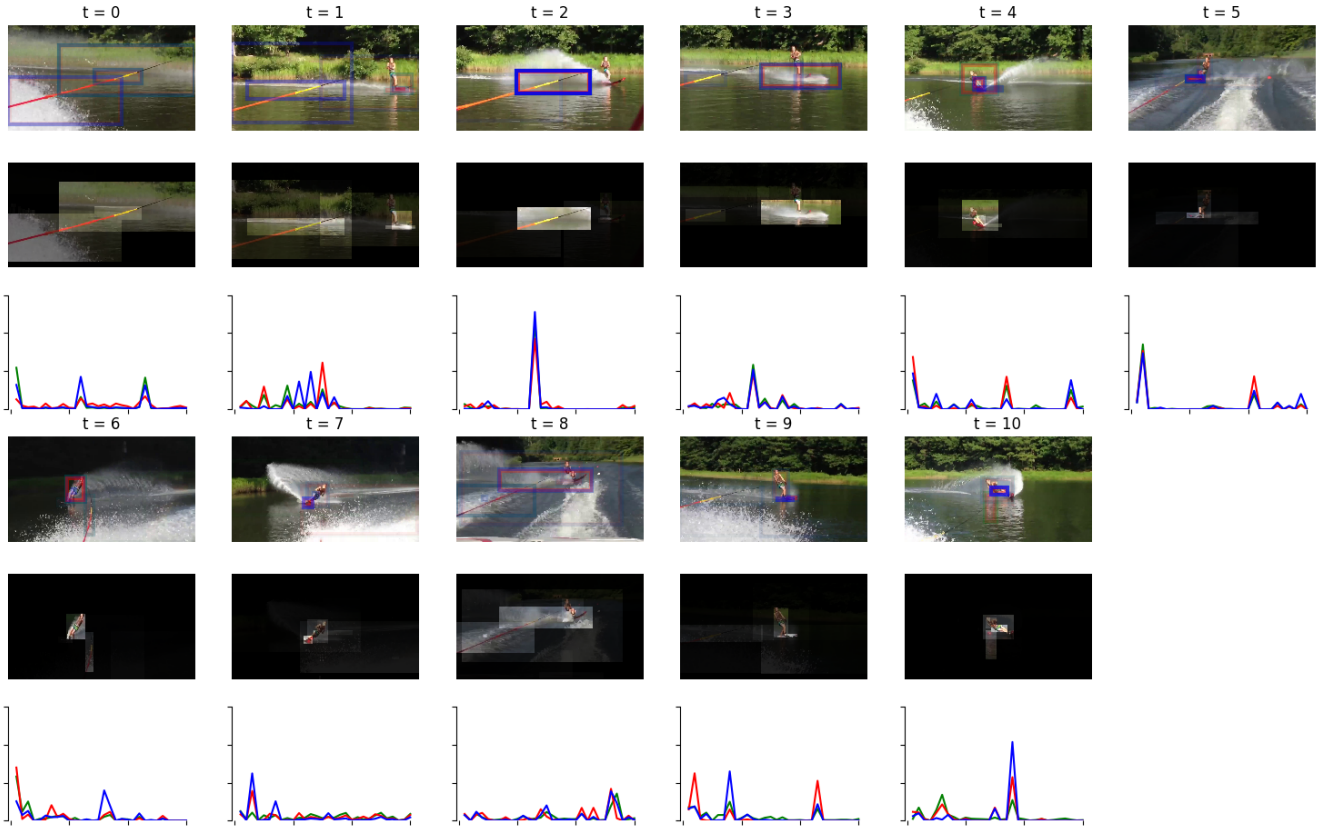

Figure 3. **Water skiing**: Our SINet is able to identify several object relationships and reasons these interactions through time: (1) the rope above the water (2) the wakeboard on the water (3) human riding on the wakeboard (4) rope connecting to the person on the wakeboard. From the distribution of three different attention weights (red, green, blue), we can also see that the proposed attention method not only is able to select objects with different inter-relationships but also can use a common object to discover different relationships around that object when needed. We observed that our method tends to explore the whole scene at the beginning of the video, and focus on new information that is different from the past. For example, while video frame at first few frames are similar, the model focus on different aspect of the visual representation.

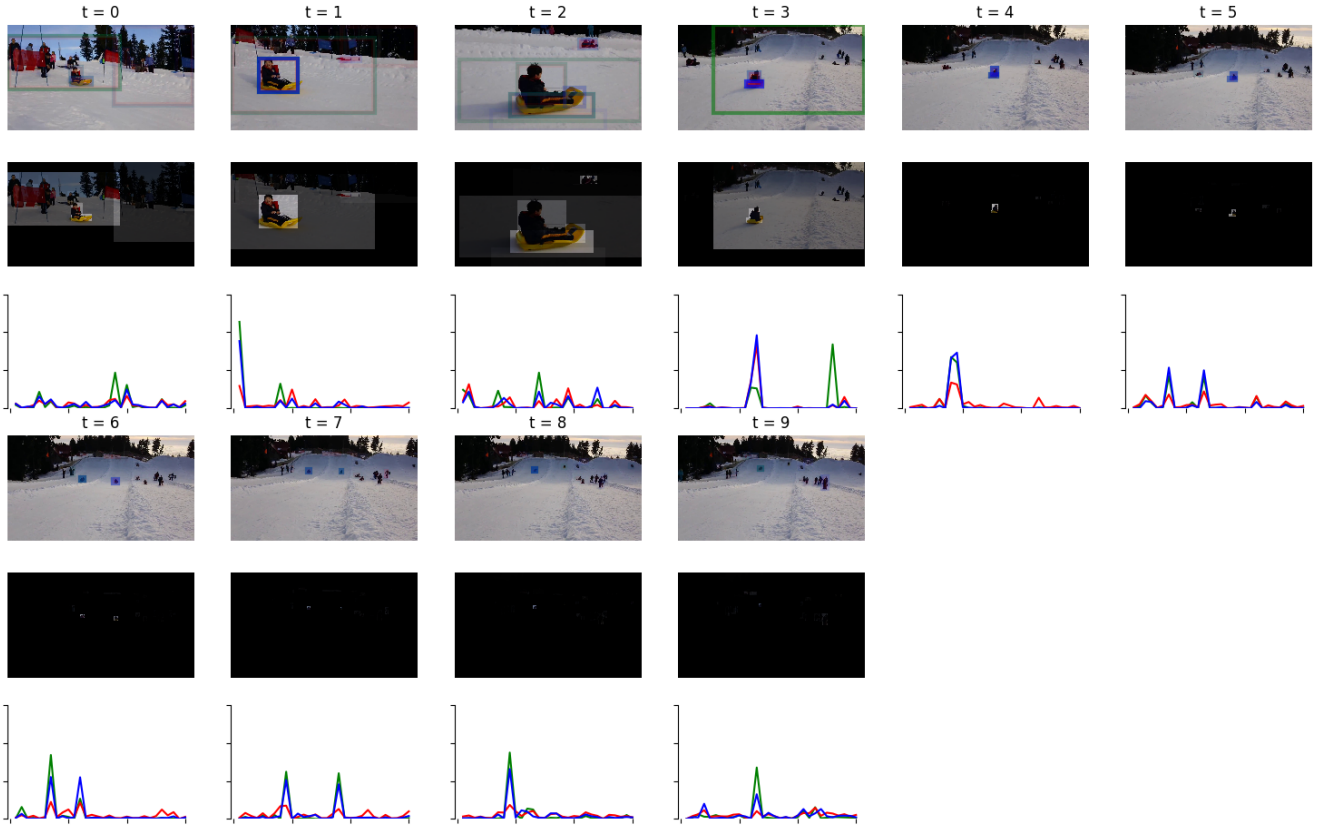

Figure 4. **Tobogganing**: Identifying *Tobogganing* essentially need three elements: toboggan, snow scene, and a human sitting on top. The three key elements are accurately identified and their interaction are highlighted as we can see from $t = 1$ to $t = 3$. Note that the model is able to continue tracking the person and toboggan throughout the whole video, even though they appear very small towards the end of the video. We can also noticed that our SINet completely ignore the background scene in the last several video frames as they are not informative since they can be easily confused by other 18 action classes involving snow and ice, e.g. *Making snowman*, *Ski jumping*, *Skiing crosscountry*, *Snowboarding*, etc.

Table 1. METEOR, ROUGE-L, CIDEr-D, and BLEU@N scores on the ActivityNet Captions 1st and 2nd validation set. All methods use ground truth temporal proposal, and out results are evaluated using the code provided in [2] with $tIoU = 0.9$. Our results with ResNeXt spatial features use videos sampled at maximum 1 FPS only.

Method	B@1	B@2	B@3	B@4	ROUGE-L	METEOR	CIDEr-D
1st Validation set							
SINet-Caption — img (C3D)	16.93	7.91	3.53	1.58	18.81	8.46	36.37
SINet-Caption — img (ResNeXt)	18.71	9.21	4.25	2.00	20.42	9.55	41.18
SINet-Caption — obj (ResNeXt)	19.00	9.42	4.29	2.03	20.61	9.50	42.20
SINet-Caption — img + obj — no co-attention (ResNeXt)	**19.89**	9.76	4.48	2.15	21.00	9.62	43.24
SINet-Caption — img + obj (ResNeXt)	19.63	**9.87**	**4.52**	**2.17**	**21.22**	**9.73**	**44.14**
2nd Validation set							
SINet-Caption — img (C3D)	17.42	8.07	3.53	1.35	18.75	8.41	40.06
SINet-Caption — img (ResNeXt)	18.91	9.41	4.28	1.68	20.49	9.56	45.05
SINet-Caption — obj (ResNeXt)	19.14	9.53	4.47	1.81	20.73	9.61	45.84
SINet-Caption — img + obj — no co-attention (ResNeXt)	**19.97**	9.88	**4.55**	**1.90**	21.15	**9.96**	**46.37**
SINet-Caption — img + obj (ResNeXt)	19.92	**9.90**	4.52	1.79	**21.28**	9.95	45.54

Figure 5. **Abseiling** is challenging since there are similar classes exist: *Climbing a rope*, *Diving cliff*, and *Rock climbing*, which involve ropes, rocks and cliffs. To achieve this, the model progressively identify the interactions and relationships like: human sitting the rock, human holding the rope, and the presence of both rope and rock. This information is proven to be sufficient for predicting *Abseiling* over other ambiguous action classes.

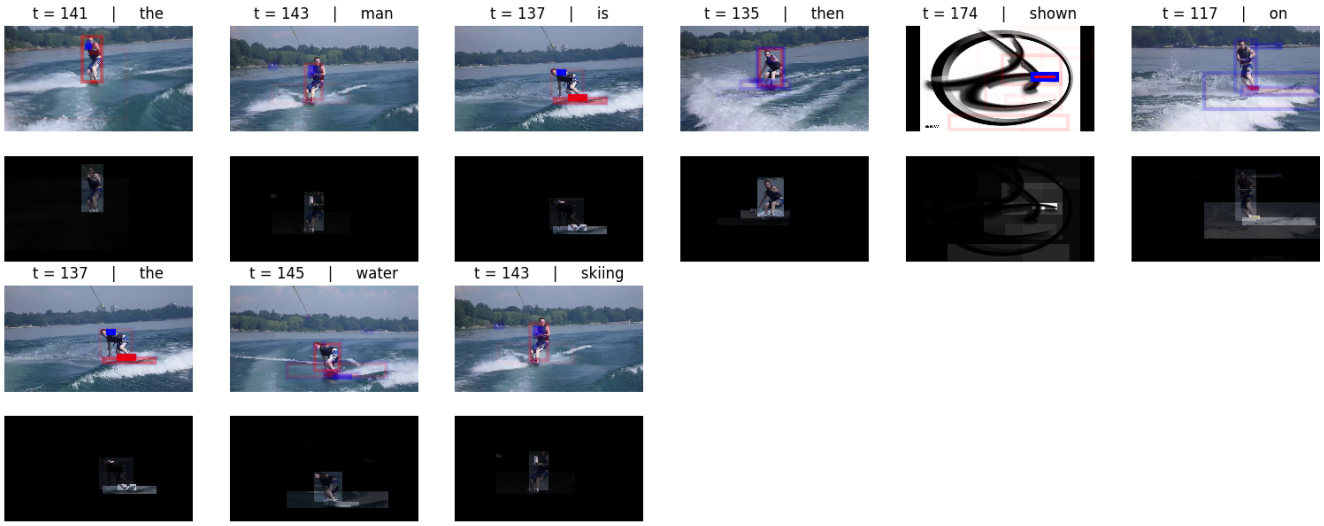

Figure 6. *The man is then shown on the water skiing.* We can see that the proposed SINet-Caption often focus on the person and the wakeboard, and most importantly it highlight the interaction between the two, i.e. the person steps on the wakeboard.

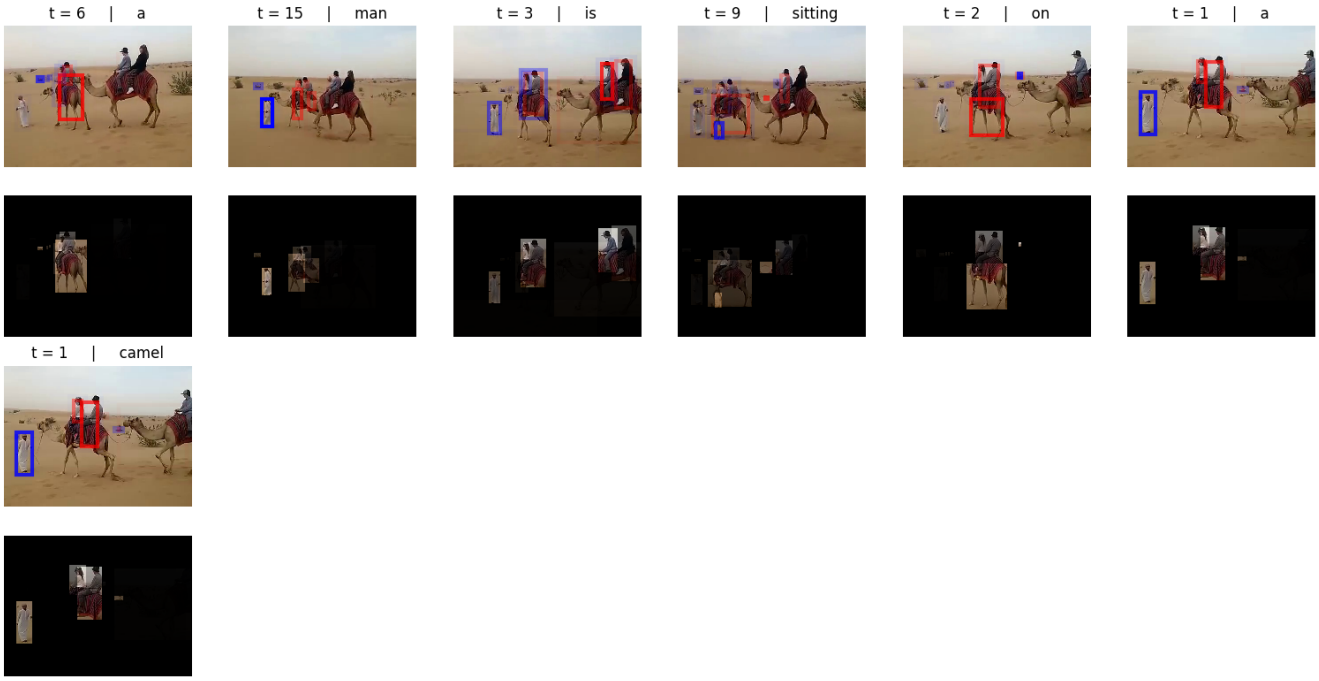

Figure 7. *A man is sitting on a camel.* The SINet-Caption is able to detect the ROIs containing both persons and the camel. We can also observe that it highlights both the ROIs for persons who sit on the camel and the camel itself at frame 3 and 9. However, the proposed method failed to identify that there are multiple people sitting on two camels. Furthermore, in some cases, it selects the person who leads the camels. This seems to be because the same video is also annotated with another caption focusing on that particular person: *A short person that is leading the camels turns around.*

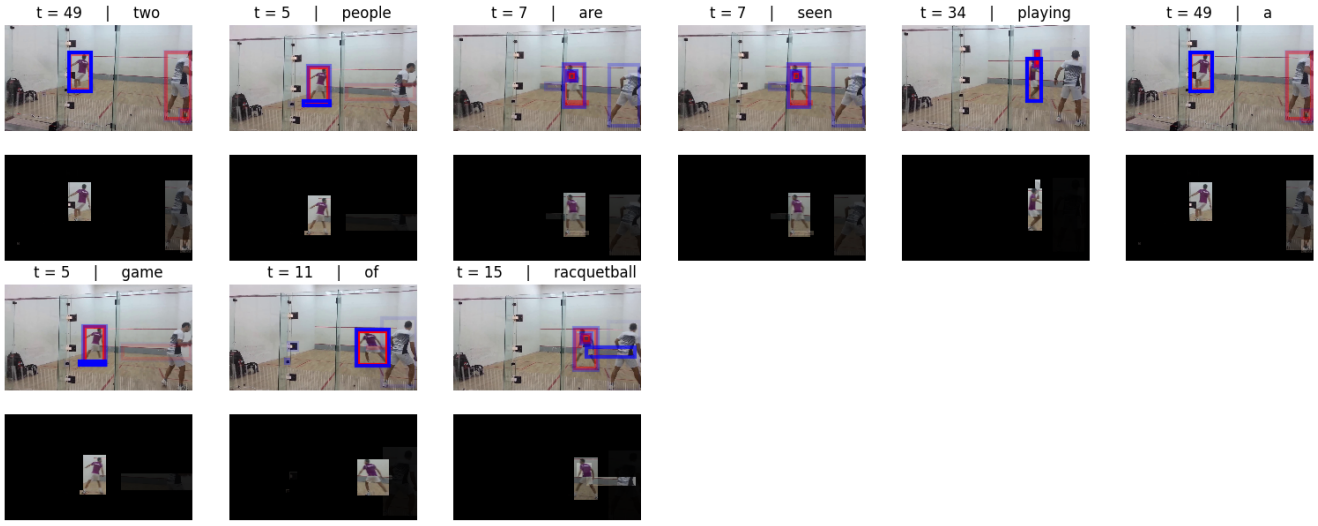

Figure 8. *Two people are seen playing a game of racquetball.* The SINet-Caption is able to identify that two persons are playing the racquetball and highlight the corresponding ROIs in the scene.

Table 2. FLOPs calculation on Kinetics sampled at 1 FPS. The calculation is based on forward passing of one video.

Proposed method ($K = 2$)		FLOP	Object pairs		FLOP
Project obj features					
MLP $g_{\theta_k(o_{i,t})}$	15 x 2048 x 2048 x 2	0.13e9	MLP	105 x 4096 x 2048	0.9e9
	15 x 2048 x 2048 x 2	0.13e9		105 x 2048 x 2048	0.4e9
	15 x 2048 x 2048 x 2	0.13e9		105 x 2048 x 2048	0.4e9
Recurrent unit					
Recurrent HOI (SDP-Attention)					
$W_h h_{t-1}$	2048 x 2048 x 2	8.4e6			
$W_c v_{c,t}$	2048 x 2048 x 2	8.4e6			
MatMul	15 x 15 x 2048 x 2	0.9e6			
MatMul	15 x 15 x 2048 x 2	0.9e6			
LSTM Cell	8 x 2 x 2 x 2048 x 2048	134.2e6	LSTM Cell	8 x 2 x 2048 x 2048	67e6
Total					
timesteps ($T = 10$)	10 x (MLP + Recurrent)	5.3e9		10 x (MLP + Recurrent)	18.3e9