Grounded Objects and Interactions for Video Captioning

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

Abstract

We address the problem of video captioning by grounding language generation on object interactions in the video. Existing work mostly focuses on overall scene understanding with often limited or no emphasis on object interactions to address the problem of video understanding. In this paper, we propose SINet-Caption that learns to generate captions grounded over higher-order interactions between arbitrary groups of objects for fine-grained video understanding. We discuss the challenges and benefits of such an approach. We further demonstrate state-of-the-art results on the ActivityNet Captions dataset using our model, SINet-Caption based on this approach.

1 Introduction

Video understanding using natural language processing to generate captions for videos has been regarded as a crucial step towards machine intelligence. However, like other video understanding tasks addressed using deep learning, initial work on video captioning has focused on learning compact representations combined over time that is used as an input to a decoder, e.g. LSTM, at the beginning or at each word generation stage to generate a caption for the target video [19, 28, 27]. This has been improved by using spatial and temporal attention mechanisms [25, 30, 32] and/or semantics attribute methods [6, 20, 24, 33]. These methods do not ground their predictions on object relationships and interactions, i.e. they largely focus on overall scene understanding or certain aspects of the video at best. However, modeling visual relationships and object interactions in a scene is a crucial form of video understanding as shown in Figure 1.

There has been considerable work that detect visual relationships in images using separate branches in a ConvNet to explicitly model object, human, and their interactions [3, 7], using scene graphs [10, 14, 15, 29] and by pairing different objects in the scene [4, 9, 23, 34]. While these models can successfully detect visual relationships for images, these methods are intractable in the video domain making it inefficient if not impossible to detect all relationships across all individual object pairs [35]. As a result, past work has at best focused on pairwise relationships on limited datasets [13, 18].

In this paper, we first hypothesize using these interactions as basis for caption generation. We then describe our method to achieve this goal by using an Region Proposal Network (RPN) to extract ROIs from videos and learning their interactions efficiently using dot product attention. Our model, **SINet-Caption** efficiently explores and grounds caption generation over interactions between arbitrary subgroups of objects, the members of which are determined by a learned attention mechanism. In our results, we demonstrate how to exploit both overall image context (coarse-grained) and higher-order object interactions (fine-grained) in the spatiotemporal domain for video captioning, as illustrated in Figure 2. We obtain state-of-the-art results on video captioning over the challenging ActivityNet Captions dataset.

ViGIL Workshop in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Figure 1: Video captions are composed of multiple visual relationships and interactions. We detect higher-order object interactions and use them as basis for video captioning.

2 Video captioning model

The proposed **SINet-Caption** first attentively models object inter-relationships and discovers the higher-order interactions for a video. The detected higher-order object interactions (fine-grained) and overall image representation (coarse-grained) are then temporally attended as visual cue for each word generation.

2.1 Higher-order object interactions

Problem statement: We define *objects* to be a certain region in the scene that might be used to determine the visual relationships and interactions. Each object representation can be directly obtained from an RPN and further encoded into an object feature, as shown in Figure 2. Note that we do not encode class information from the object detector since there exists cross-domain problem, and we may miss some objects that are not detected by pre-trained object detector. Also, we do not know the corresponding object across time since linking objects through time may not be suitable if the video sequence is long and computationally expensive. Our objective is to efficiently detect higher-order interactions—*interactions beyond pairwise objects*—from these rich yet unordered object representations reside in a high-dimensional space that spans across time.

In the simplest setting, an interaction between objects in the scene can be represented as combining individual object information. For example, one method is to add the learnable representations and project these representations into a high-dimensional space where the object interactions can be exploited by simply summing up the object representations [23]. Another approach which has been widely used with images is pairing all possible object candidates (or subject-object pairs) [3, 4, 9, 34]. However, this is infeasible for video, since a video typically contains hundreds of frame and the set of object-object pairs is too large to fully represent. Detecting object relationships frame by frame is computationally expensive, and the temporal reasoning of object interactions is not used.

Recurrent Higher-Order Interaction Module: To overcome these issues, we propose a generic recurrent module for detecting higher-order object interactions for fine-grained video understanding problem, as illustrated in Figure 3 (right). The proposed recurrent module dynamically selects object candidates which are important to describe video content. The combinations of these selected objects are then concatenated to model higher-order interaction using group to group or triplet groups of objects.

Formally, the current image representation $v_{c,t}$ and previous object interaction representation h_{t-1} are first projected to introduce learnable weights. The projected $v_{c,t}$ and h_{t-1} are then repeated and expanded n times (the number of objects at time t). We directly combine this information with projected objects via matrix addition and use it as input to dot-product attention. In short, the attention

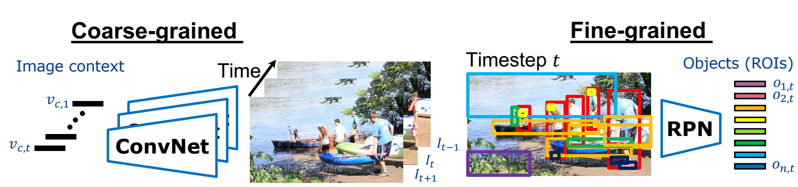

Figure 2: We exploit both coarse- (overall image) and fine-grained (object) visual representations for each video frame.

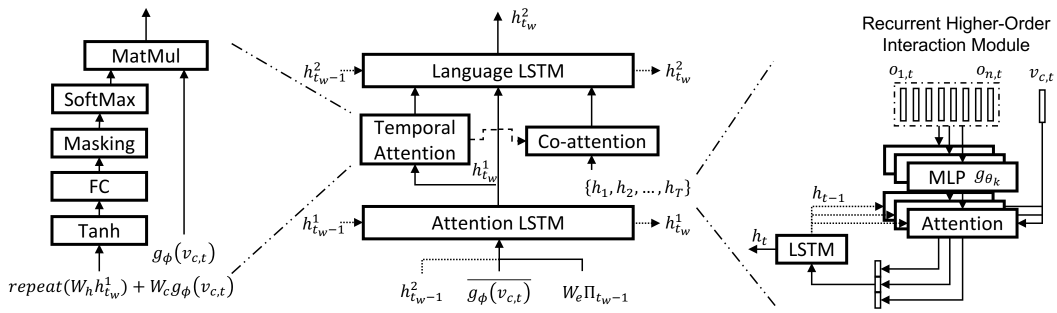

Figure 3: Overview of the proposed SINet-Caption for video captioning.

is computed using inputs from current (projected) object features, overall image visual representation, and previously discovered object interactions, which provide the attention mechanism with maximum context. Specifically, the input to the attention can be define as: $X_k = repeat(W_{h_k} h_{t-1} + W_{c_k} v_{c,t}) + g_{\theta_k}(O_t)$, and the attention weights over all objects can be define as: $\alpha_k = softmax(\frac{X_k{}^\top X_k}{\sqrt{d_\theta}})$, where W_{h_k} and W_{c_k} are learned weights for h_{t-1} and $v_{c,t}$. g_{θ_k} is a Multi-Layer Perceptron (MLP) with parameter θ_k, d_θ is the dimension of last fully-connected layer of g_{θ_k}, X_k is the input to kth attention module, and $\sqrt{d_\theta}$ is a scaling factor. We omit the bias term for simplicity.

The attended object feature at time t is then calculated as mean-pooling on weighted objects: $v^k_{o,t} = \overline{\alpha_k \; (g_{\theta_k}(O_t))^\top}$. The output $v^k_{o,t}$ is a single feature vector representation which encodes the kth object inter-relationships of a video frame at time t, and k ranges from $\{1, 2, ..., K\}$ representing the number of groups for inter-relationships.

Finally, the selected object candidates $v^k_{o,t}$ are then concatenated and used as the input to a LSTM cell: $h_t = LSTM(v^1_{o,t} \| v^2_{o,t} \| ... \| v^K_{o,t})$. The output h_t is then defined as the higher-order object interaction representation at current time t. The sequence of hidden state of the LSTM cell $\mathbf{h} = (h_1, h_2, ..., h_T)$ are the representations of higher-order object interactions for each timestep.

2.2 Video captioning with coarse- and fine-grained

We now describe how our **SINet-Caption** exploit both coarse-grained (overall image representation) and fine-grained (higher-order object interactions) for video captioning.

The SINet-Caption is inspired by prior work using hierarchical LSTM for captioning task [1, 25], and we extend it with the proposed recurrent higher-order interaction module so that the model can leverage the detected higher-order object interactions, as shown in Figure 3. The model consist of two LSTM layers: Attention LSTM and Language LSTM.

Attention LSTM: The Attention LSTM fuses the previous hidden state output of Language LSTM $h^2_{t_w-1}$, overall representation of the video, and the input word at time $t_w - 1$ to generate the hidden representation for the following attention module. Formally, the input to Attention LSTM can be defined as: $x^1_{t_w} = h^2_{t_w-1} \; \| \; \overline{g_\phi(v_{c,t})} \; \| \; W_e \Pi_{t_w-1}$, where $\overline{g_\phi(v_{c,t})}$ is the projected and mean-pooled image features, g_ϕ is a MLP with parameter ϕ, $W_e \in \mathbb{R}^{E \times \Sigma}$ is a word embedding matrix for a vocabulary of size Σ, and Π_{t_w-1} is one-hot encoding of the input word at time $t_w - 1$. Note that t is the video time, and t_w is the timestep for caption generation.

Temporal attention: The input for computing the temporal attention is the combination of output of the Attention LSTM $h^1_{t_w}$ and projected image features $g_\phi(v_{c,t})$, and the attention is computed using a simple *tanh* function and a fully-connected layer to attend to projected image features $g_\phi(v_{c,t})$, as illustrated in Figure 3. Specifically, the input can be defined as $x_a = repeat(W_h h^1_{t_w}) + W_c g_\phi(v_{c,t})$, and the temporal attention can be obtained by $\alpha_{temp} = softmax(w_a{}^\top tanh(x_a))$, where $W_h \in \mathbb{R}^{d_{g_\phi} \times d_h}$ and $W_c \in \mathbb{R}^{d_{g_\phi} \times d_{g_\phi}}$ are learned weights for h^1_t and $g_\phi(v_{c,t})$. d_{g_ϕ} is the dimension of last fully-connected layer of g_ϕ.

Co-attention: We directly apply the temporal attention obtained from image features on object interaction representations $\mathbf{h} = (h_1, h_2, ..., h_T)$ (see Sec 2.1 for details).

Table 1: METEOR [2], ROUGE-L [16], CIDEr-D [26], and BLEU@N [21] scores on the ActivityNet Captions test and validation set. All methods use ground truth proposal except LSTM-A_3 [31]. Our results with ResNeXt spatial features use videos sampled at maximum 1 FPS only.

Method	B@1	B@2	B@3	B@4	R	M	C
Test set							
LSTM-YT [27] (C3D)	18.22	7.43	3.24	1.24	-	6.56	14.86
S2VT [28] (C3D)	20.35	8.99	4.60	2.62	-	7.85	20.97
H-RNN [32] (C3D)	19.46	8.78	4.34	2.53	-	8.02	20.18
S2VT + full context [12] (C3D)	26.45	13.48	7.21	3.98	-	9.46	24.56
LSTM-A_3 + policy gradient + retrieval [31] (ResNet + P3D ResNet [22])	-	-	-	-	-	12.84	-
Validation set (Avg. 1st and 2nd)							
LSTM-A_3 (ResNet + P3D ResNet) [31]	-	-	-	**3.38**	13.27	7.71	16.08
LSTM-A_3 + policy gradient + retrieval [31] (ResNet + P3D ResNet [22])	-	-	-	3.13	14.29	8.73	14.75
SINet-Caption | img (C3D)	17.18	7.99	3.53	1.47	18.78	8.44	38.22
SINet-Caption | img (ResNeXt)	18.81	9.31	4.27	1.84	20.46	9.56	43.12
SINet-Caption | obj (ResNeXt)	19.07	9.48	4.38	1.92	20.67	9.56	44.02
SINet-Caption | img+obj | no co-attn (ResNeXt)	**19.93**	9.82	**4.52**	2.03	21.08	9.79	44.81
SINet-Caption | img+obj (ResNeXt)	19.78	**9.89**	**4.52**	1.98	**21.25**	**9.84**	**44.84**

Language LSTM: Finally, the Language LSTM takes the concatenation of output of the Attention LSTM $h^1_{t_w}$, attended video representation $\hat{v}_{c,t_w}$, and co-attended object interactions $\hat{h}_{t_w}$ as input: $x^2_{t_w} = h^1_{t_w} \parallel \hat{v}_{c,t_w} \parallel \hat{h}_{t_w}$. The output of Language LSTM is then used to generate words via fully-connected and softmax layer.

3 Evaluation on ActivityNet Captions

We use the ActivityNet Captions dataset for evaluating SINet-Caption. The ground truth temporal proposals are used to segment videos, i.e. we treat each video segment independently since our focus in this work is modeling object interactions for video captioning rather than on temporal proposals (please refer to supplementary material for details on dataset and implementation). All methods in Table 1 use ground truth temporal proposal, except LSTM-A_3 [31].

For a fair comparison with prior methods that use C3D features, we report results with both C3D and ResNeXt spatial features. Since there is no prior result reported on the validation set, we compare our method via LSTM-A_3 [31] which reports results on the validation and test sets. This allows us to indirectly compare with methods reported on the test set. As shown in Table 1, while LSTM-A_3 clearly outperforms other methods on the test set with a large margin, our method shows better results on the validation sets across nearly all language metrics. Note that we do not claim our method to be superior to LSTM-A_3 because of two fundamental differences. First, they do not rely on ground truth temporal proposals. Second, they use features extracted from a ResNet fine-tuned on Kinetics and another P3D ResNet [22] fine-tuned on Sports-1M, whereas we only use an ResNeXt-101 fine-tuned on Kinetics sampled at maximum 1 FPS. Utilizing more powerful feature representations can improve the prediction accuracy by a large margin on video captioning tasks. This also corresponds to our experiments with C3D and ResNeXt features, where the proposed method with ResNeXt features performs significantly better than C3D features. We observed that using only the detected object interactions shows slightly better performance than using only overall image representation. This demonstrates that even though the SINet-Caption is not aware of the overall scene representation, it achieves similar performance relying on only the detected object interactions. By combining both, the performance further improved across all evaluation metrics, with or without co-attention.

In conclusion, we introduce a generic recurrent module to detect higher-order object interactions for video understanding task. We demonstrate on ActivityNet Captions that the proposed SINet-Caption exploiting both higher-order object interactions (fine-grained) and overall image representation (coarse-grained) outperforms prior work by a substantial margin.

Supplementary Materials

Dataset and Implementation

ActivityNet Captions dataset: We use the new ActivityNet Captions for evaluating SINet-Caption. The ActivityNet Captions dataset contains 20k videos and has total 849 video hours with 100k total descriptions. We focus on providing the fine-grained understanding of the video to describe video events with natural language, as opposed to identifying the temporal proposals. We thus use the ground truth temporal segments and treat each temporal segment independently. We use this dataset over other captioning datasets because ActivityNet Captions is action-centric, as opposed to object-centric [12]. This fits our goal of detecting higher-order object interactions for understanding human actions. Following the same procedure as in [12], all sentences are capped to be a maximum length of 30 words. We sample predictions using beam search of size 5 for captioning. While the previous work sample C3D features every 8 frames [12], we only sampled video at maximum 1 FPS. Video segments longer than 30 seconds are evenly sampled at maximum 30 samples.

Image feature: We pre-trained a ResNeXt-101 on the Kinetics dataset [11] sampled at 1 FPS (approximately 2.5 million images), and use it to extract image features. We use SGD with Nesterov momentum as the optimizer. The initial learning rate is $1e-4$ and automatically drops by 10x when validation loss saturated for 5 epochs. The weight decay is $1e-4$ and the momentum is 0.9, and the batch size is 128. We follow the standard data augmentation procedure by randomly cropping and horizontally flipping the video frames during training. When extracting image features, the smaller edge of the image is scaled to 256 pixels and we crop the center of the image as input to the fine-tuned ResNeXt-101. Each image feature is encoded to a 2048-d feature vector.

Object feature: We generate the object features by first obtaining the coordinates of ROIs from a Deformable R-FCN [5] with ResNet-101 [8] as the backbone architecture. The Deformable R-FCN was trained on MS-COCO train and validation dataset [17]. We set the IoU threshold for NMS to be 0.2. For each of the ROIs, we extract features using ROI coordinates and adaptive max-pooling from the same ResNeXt-101 model pre-trained on Kinetics. The resulting object feature for each ROI is a 2048-d feature vector. ROIs are ranked according to their ROI scores. We select top 15 objects only. We do not use object class information since there is a cross-domain problem and we may miss some objects that were not detected. For the same reason, the bounding-box regression process is not performed here since we do not have the ground-truth bounding boxes.

Training: We train the proposed SINet-Caption with ADAM optimizer. The initial learning rate is set to $1e-3$ and automatically drops by 10x when validation loss is saturated. The batch sizes is 32.

ActivityNet Captions on 1st and 2nd validation set

We report the performance of SINet-Caption on both 1st and 2nd validation set in Table 2.

Table 2: 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 [12] 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	R	M	C
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-attn (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-attn (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 4: *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.

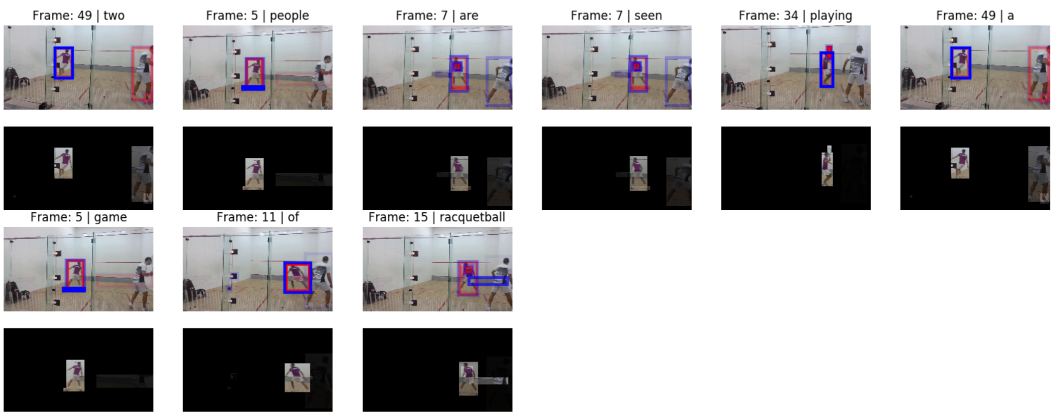

Figure 5: *Two people are seen playing a game of racquetball.* The SINet-Caption is able to identify that two persons are playing the racquetball and highlight the corresponding ROIs in the scene.

Qualitative Analysis

To further validate the proposed method, we qualitatively show how the SINet-Caption selectively attends to various video frames (temporal), objects, and object relationships (spatial) during each word generation. Several examples are shown in Figure 4, 5, 6, and 7. Note that 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). Also, since we aggregate the detected object interactions via 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.

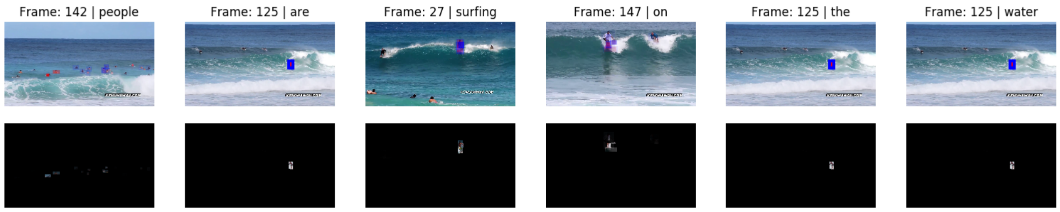

Figure 6: *People are surfing on the water.* At the beginning, the SINet-Caption identify multiple people are in the ocean. The person who is surfing on the water is then successfully identified, and the rest of irrelevant objects and background are completely ignored.

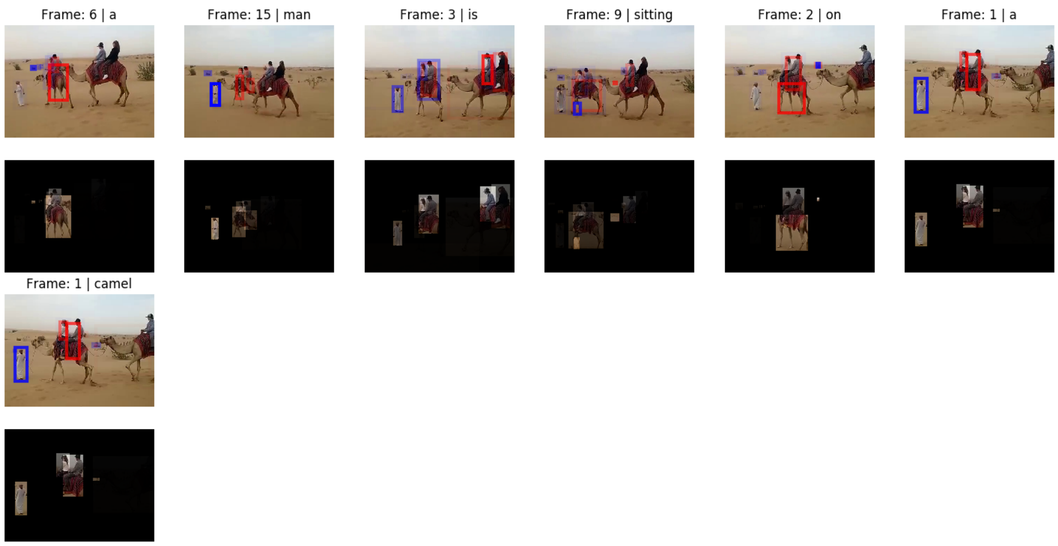

Figure 7: *A man is sitting on a camel.* The SINet-Caption is able to detect the ROIs containing both persons and the camel. We can also observe that it highlights both the ROIs for persons who sit on the camel and the camel itself at frame 3 and 9. However, the proposed method failed to identify that 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.*

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.

[3] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. *arXiv preprint arXiv:1702.05448*, 2017.

[4] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. *arXiv preprint arXiv:1704.03114*, 2017.

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *arXiv preprint arXiv:1703.06211*, 2017.

[6] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. *arXiv preprint arXiv:1611.08002*, 2016.

[7] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. *arXiv preprint arXiv:1704.07333*, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[9] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. *arXiv preprint arXiv:1611.09978*, 2016.

[10] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015.

[11] 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.

[12] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. *arXiv preprint arXiv:1705.00754*, 2017.

[13] C. Lea, A. Reiter, R. Vidal, and G. D. Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2016.

[14] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1261–1270, 2017.

[15] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *arXiv preprint arXiv:1703.03054*, 2017.

[16] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] B. Ni, V. R. Paramathayalan, and P. Moulin. Multiple granularity analysis for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 756–763, 2014.

[19] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016.

[20] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. *arXiv preprint arXiv:1611.07675*, 2016.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[22] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017.

[24] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue. Weakly supervised dense video captioning. *arXiv preprint arXiv:1704.01502*, 2017.

[25] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, and H. T. Shen. Hierarchical lstm with adjusted temporal attention for video captioning. *arXiv preprint arXiv:1706.01231*, 2017.

[26] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[27] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[28] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[29] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. *arXiv preprint arXiv:1701.02426*, 2017.

[30] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.

[31] T. Yao, Y. Li, Z. Qiu, F. Long, Y. Pan, D. Li, and T. Mei. Activitynet challenge 2017: Dense-captioning events in videos. https://drive.google.com/file/d/0BxMIVlhgRmcbd0VVUEE4RDhPV2s/, 2017.

[32] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016.

[33] Y. Yu, H. Ko, J. Choi, and G. Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, volume 3, page 7, 2017.

[34] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. *arXiv preprint arXiv:1702.08319*, 2017.

[35] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal. Relationship proposal networks. In *CVPR*, volume 1, page 2, 2017.