INFORMATIVE ATTENTION SUPERVISION FOR GROUNDED VIDEO DESCRIPTION

Boyang Wan, Wenhui Jiang, Yuming Fang,

Jiangxi University of Finance and Economics, School of Information Management, Nanchang, China

ABSTRACT

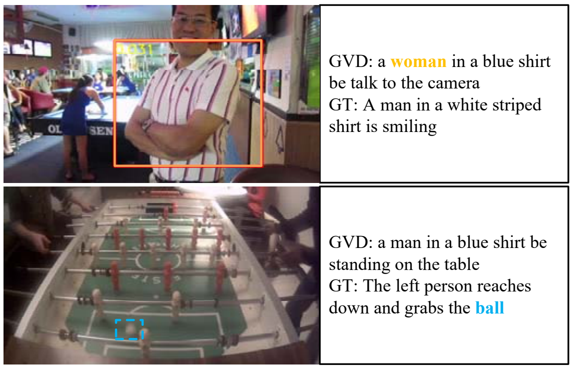
Attention supervision encourages grounded video description models (GVDMs) to focus on the related visual content when generating words. Thus, it improves the description performance of GVDMs. However, existing GVDMs often fail to focus on small but informative regions because these regions are considered as negative by using the intersection-over-union (IoU) based attention groundtruth sampling method. Moreover, the prevailing attention loss functions enforce the GVDMs to focus equally on all sampled regions when the GVDMs generate words, which may make it difficult for the model to attend to informative regions and thus degrade the quality of the generated sentences. To alleviate the above problems, we propose an informative attention supervision method including a novel attention groundtruth sampling method and a group-based weak grounding supervision. Specifically, our attention groundtruth sample method captures small proposal regions that overlap with the entity boxes. The proposed grounding supervision allows the GVDMs to dynamically focus on some of the most informative attention regions instead of all of them. Our approach yields competitive results on the ActivityNet Entities dataset without bells and whistles, surpassing previous methods without increasing inference costs.

Index Terms— Grounded Video Description, Attention Supervision

1. INTRODUCTION

Visual description is an essential and representative cross-media task that automatically generates sentences for describing videos or images. Recently, the visual description task gains significant development based on the large-scale visual description datasets [1, 2] and deep-learning based techniques [3, 4]. However, some studies [5, 6] reveal that prevalent visual description models [7, 2] (VDMs) suffer from object hallucination [8] and gender bias [9] while obtaining excellent performance on the large-scale visual description datasets. According to [5, 6], object hallucination and gender bias are generated by the VDMs not being well-grounded, *i.e.*, the most attended region generated by the VDMs is not aligned with the corresponding region that the words/phrases describe in the image/video.

Many studies [10, 5] have been published to improve the performance of the VDMs by using grounding methods. In [10], description sentences are generated by filling the visual concepts obtained using an object detection network into templates with slot locations. Consequently, this model is limited in the diversity of the generated sentences. On the other hand, attention supervision, which forces the VDMs to model the relationship between words and visual regions, can alleviate object hallucination and gender bias of the VDMs. Fundamentally, [5] proposes the ActivityNet Entities (AE) dataset, which provides word-regions groundtruth, to model the relationship between words and visual regions explicitly. Therefore, the grounded video description (GVD) model proposed in [5] gains improved video description performance under end-to-end language and attention supervision in a fully supervised manner.



GVD: a **woman** in a blue shirt be talk to the camera
GT: A man in a white striped shirt is smiling

GVD: a man in a blue shirt be standing on the table
GT: The left person reaches down and grabs the **ball**

Fig. 1. Failure examples of description and attention results achieved by the GVD [5]. The solid line box and the number in solid line box represent the most attended proposal region and the attention weight for generated noun in color and bold, respectively. The dotted box is denoted a nouns in color and bold from groundtruth sentence.

However, the GVD often fails to focus on small but informative regions, predicting incorrect nouns or missing visual objects, as shown in Fig. 1. Compared to 'head', which occupies a relatively small region in the frames and dominates gender judgment, GVD focuses more on the 'body' region. Such undesirable behavior is mainly caused by two reasons. Firstly, GVD tends to focus on object parts with large intersection-over-union (IoU) with target regions. It may be caused by small regions not being sampled as attended regions based on their small IoU with target regions, even though they may be informative regions. Secondly, the entropy-based attention loss function used in [5] enforces GVD to attend all positive regions simultaneously. As a result, GVD fails to concentrate on the informative part of positive regions, which deteriorates the captioning performance considerably.

To address the above issues, we build a new attention supervision method, including a proposal based attention groundtruth sampling method and a group based weak grounding supervision. The proposal based attention groundtruth sampling method provides smaller regions of

This work has been supported by the National Natural Science Foundation of China (No.62161013 and No.62132006), the Key Research and Development Program of Jiangxi Province (No.20203BBE53033 and No.20202ACB202007), the Science and Technology Innovation Commission of Shenzhen (No.2021Szvup051) and the 16th student scientific research project of Jiangxi University of Finance and Economics(No.20210918115557066). The source code will provide in https://github.com/wanboyang/IASGV_ICASSP2022.

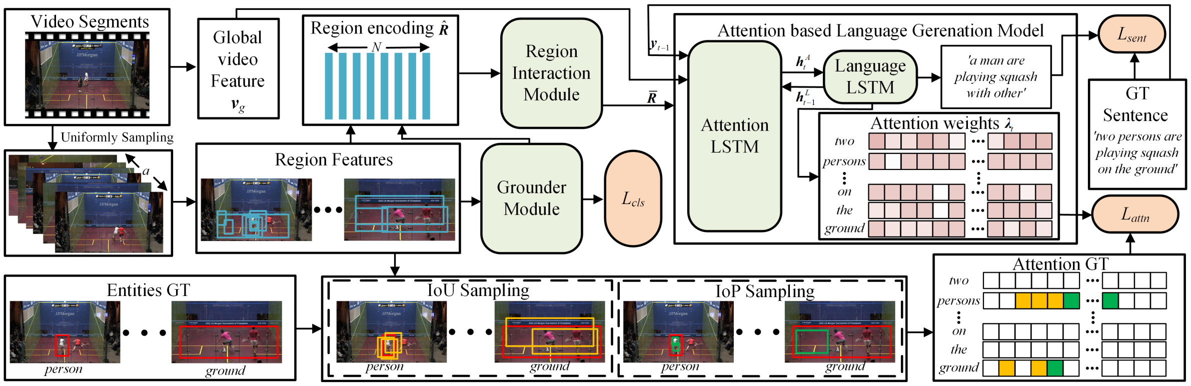


Fig. 2. The proposed framework. Red boxes and blue boxes indicate entity annotations and proposal regions, and yellow boxes and green boxes denote IoU-based and IoP-based sampled proposal regions, respectively.

objects. Motivated by Multiple-Instance Learning (MIL) techniques [11, 12], the group based weak grounding supervision implemented by an attention loss function treats the attention groundtruth as the weak supervision. It forces the model to dynamically attend partial but informative regions, rather than all sampled regions. Thus, compared to GVD, our method alleviates the issue of neglected informative regions and guides video description models to generate more precise nouns in descriptions.

In summary, our main contributions are listed as follows:

- The first attempt to explore a universal framework to utilize the small but informative regions for grounded video description task.
- A novel attention groundtruth sampling method and the corresponding weak grounding supervision method.
- Achieving a new SOTA performance on grounded video description task.

2. PROPOSED METHOD

In this section, we first define the notation and baseline. Then we describe the proposed informative attention supervision method. Specifically, we present our proposal based attention groundtruth sampling method followed by a detailed description of the group based weak grounding supervision for the baseline model.

2.1. Notation

The input video segment and the sentence are denoted as $\boldsymbol{V}$ and $\boldsymbol{S}$, respectively. We uniformly sample a frames from each input video segment and the sampled frames are represented as $\boldsymbol{F} = \{\boldsymbol{F}_i\}_{i=1}^{a}$. Further, a set of regions, which is obtained by using a pretrained object detection model, is extracted from $\boldsymbol{F}$, and the embedding of regions are denoted as $\boldsymbol{R} = \{\boldsymbol{r}_i\}_{i=1}^{N} \in \mathbb{R}^{d \times N}$, where d and N are feature dimension and number of regions. The words in sentences $\boldsymbol{S}$ are denoted as $\{\boldsymbol{s}_i\}_t$, respectively, where $t \in \{1, 2, .., T\}$. The T indicates the sentence length and the one-hot embedding of the words are $\boldsymbol{y}_t \in \mathbf{R}^e$ where e is the embedding size of words.

2.2. Overall Framework

To validate the proposed informative attention supervision methods, we introduce the video event description model **GVD** from [5], which contains the grounder module, region interaction module and attention based language generation model as our baseline.

The framework of our method is presented in Fig. 2. First, a video segment is uniformly extracted with a frames. Then we get spatio-temporal coordinates and representations of the proposal regions of each frame by using an object detection network [13], which is pretrained by using Visual Genome (VG) [14], as the region feature of this video segment. The grounder module is used to determine the class of proposal regions and can be formulated as follows:

$$M_s(\boldsymbol{R}) = \mathrm{Softmax}(\boldsymbol{W}_c^{\top}\boldsymbol{R} + \boldsymbol{B}) \tag{1}$$

where $\boldsymbol{M}_s$ is the region-class similarity matrix of proposal regions $\boldsymbol{R}$, $\boldsymbol{W}_c$ and $\boldsymbol{B}$ are learnable weights and bias, respectively. The Softmax operator ensures that the sum of class distribution is 1 for each proposal region. It should note that the class label of proposal regions is the nearest neighbor of their VG classes in the AE classes. The distance of the VG classes and AE classes are obtained from the class word's glove vector [15].

To utilize the spatio-temporal cue of proposal regions, the spatial and temporal coordinates of each proposal region are projected to the d_l location embedding $\boldsymbol{M}_l$. Eventually, the region feature, the region-class similarity matrix and the location embedding are concatenated and projected into region encoding $\hat{\boldsymbol{R}} = [\hat{\boldsymbol{r}}_{\mathbf{1}}, \hat{\boldsymbol{r}}_{\mathbf{2}}, ..., \hat{\boldsymbol{r}}_{\boldsymbol{N}}]$ and the region encoding can be formulated as follows:

$$\hat{\boldsymbol{R}} = \boldsymbol{W}_g[\boldsymbol{R}|M_s(\boldsymbol{R})|\boldsymbol{M}_l] \tag{2}$$

where $[\cdot|\cdot]$ denotes a concatenation operator and $\boldsymbol{W}_g$ is fully-connected layer. Next, $\hat{\boldsymbol{R}}$ is fed into the region interaction module to get the final region encoding $\overline{\boldsymbol{R}}$. The region interaction module is a fully self-attention model that models the relationship between proposal regions. The final region encoding is provided to the attention based Language generation model to generate the caption of the input segment.

Attention based Language Generation Model. In this paper, we adopt the LSTM based model, including Attention LSTM and Language LSTM, proposed in [5] as an attention based language generation model. As illustrated in Fig. 2, the Attention LSTM is fed into the global video feature $\boldsymbol{v}_g$, previous hidden state of the Language LSTM $\boldsymbol{h}^L_{t-1}$ and previous word embedding $\boldsymbol{y}_{t-1}$ and output the hidden state $\boldsymbol{h}^A_t$. The Attention LSTM can be formulated as follows:

$$\boldsymbol{h}^A_t = \mathrm{LSTM}_{\mathrm{attn}}([\boldsymbol{v}_g|\boldsymbol{h}^L_{t-1}|\boldsymbol{y}_{t-1}]) \tag{3}$$

where $[\cdot|\cdot]$ denotes a concatenation operator. To generate a word $\boldsymbol{s}_t$, the Language LSTM uses $\boldsymbol{h}^A_t$ to generate attention weights $\boldsymbol{\lambda}_t$ for the final region encoding $\overline{\boldsymbol{R}}$,

$$\boldsymbol{\lambda}_t = \mathrm{Softmax}(\boldsymbol{W}_\lambda \tanh(\boldsymbol{W}_h\boldsymbol{h}^A_t + \boldsymbol{W}_r\overline{\boldsymbol{R}})) \tag{4}$$

$$p(\boldsymbol{s}_t) = \mathrm{Softmax}((\boldsymbol{W}_o \mathrm{LSTM}_{\mathrm{lang}}(\boldsymbol{\lambda}_t\overline{\boldsymbol{R}}, \boldsymbol{h}^A_t)) \tag{5}$$

where $\boldsymbol{W}_o$, $\boldsymbol{W}_\lambda$, $\boldsymbol{W}_h$ and $\boldsymbol{W}_r$ are learned parameters, and the $p(\boldsymbol{s}_t)$ is conditional probability distribution of $\boldsymbol{s}_t$.

2.3. Informative Attention Supervision Method

Proposal based attention groundtruth sampling method. Explicit attention supervision is a common way, which forces models to focus on attended regions, to improve the performance and interpretability of the models. Generally, the attention regions are obtained according to the intersection-over-union (IoU) of the entity boxes and the proposal regions, as shown in Fig. 2. To capture the small but informative proposal region, we propose a proposal based attention groundtruth sampling method by intersection-over-proposal (IoP) between the proposal region boxes and the annotated grounding entity boxes and can be represented as follows:

$$\mathrm{IoP}(\boldsymbol{p}, \boldsymbol{g}) = \begin{cases} \frac{I_w \cdot I_h}{w_g \cdot h_g} & I_w > 0, I_h > 0 \\ 0 & else \end{cases} \tag{6}$$

$$\begin{cases} I_l = \max(x_d, x_g) \\ I_r = \min(x_p + w_p, x_g + w_g) \\ I_t = \max(y_p, y_g) \\ I_b = \min(y_p + h_p, y_g + h_g) \\ I_w = I_r - I_l \\ I_h = I_b - I_t \\ I_s = (w_d \cdot h_d + w_g \cdot h_g) \end{cases} \tag{7}$$

where $\boldsymbol{p} = [y_p, x_p, y_p + h_p, x_p + w_p]$ and $\boldsymbol{g} = [y_g, x_g, y_g + h_g, x_g + w_g]$ represent the proposal region boxes and the annotated grounding entity boxes, respectively. Note that (y, x), h and w denote top-left point, height, and width of the input boxes. In this paper, we get attended proposal regions based on the IoU and the IoP, as shown in Fig. 2. Specifically, a proposal region can be selected as the attended proposal region when it satisfies one of the following criteria: 1.) $\mathrm{IoU}(\boldsymbol{d}, \boldsymbol{g}) >= \tau_1$, 2.) $\mathrm{IoP}(\boldsymbol{d}, \boldsymbol{g}) >= \tau_2$. The attended proposal regions seem as positive regions and the rest of the proposal regions are taken as negative regions for attention supervision.

Table 1. Comparison of ablation study on the ActivityNet Entities validation set.

$\hat{L}_{attn}$	L_{attn}	L_{cls}	IoU	IoP	B@1	B@4	M	C
✓	-	✓	✓	-	23.9	2.59	11.2	47.5
✓	-	✓	✓	✓	23.8	2.67	11.1	47.2
-	✓	✓	✓	-	23.7	2.59	11.1	47.0
-	✓	✓	✓	✓	**24.0**	**2.78**	**11.2**	**48.7**

Group based Weak Grounding Supervision. The cross-entropy based attention loss function $\hat{L}_{attn}$ proposed in [5] demands that the model focus on all the positive regions simultaneously. This may lead to ambiguity in the model's attention and a lack of focus on informative regions. To alleviate the above problems, we propose the group based weak grounding supervision that is implemented by a MIL based attention supervision function L_{attn}. Specifically, we select the regions with top-k attention weights from the positive and the negative regions to positive group G_p and negative group G_n, respectively. The model is forced to focus on the G_p as a whole and attention weights of G_p are dynamically aligned based on learning. All attention weights in G_n are forced to be 0 because negative regions do not need to be focused. Formally, L_{attn} on per-word can be represented as follows:

$$L_{attn} = -\log(\sum_{\lambda_i \in G_p} \lambda_i) + \frac{1}{|G_n|} \sum_{\lambda_i \in G_n} (\log \lambda_i) \tag{8}$$

2.4. Optimization

Following [5], we use the cross-entropy language loss L_{sent} and the region classification loss L_{cls} in the proposed model. Combining our attention loss function, the loss can be formulated as follows:

$$L = L_{sent} + \lambda_\alpha L_{attn} + \lambda_\beta L_{cls} \tag{9}$$

For a fair comparison, we set λ_β=0.1 like [5]. And we empirically set λ_α = 0.1. More details are provided in Sec.3.2.

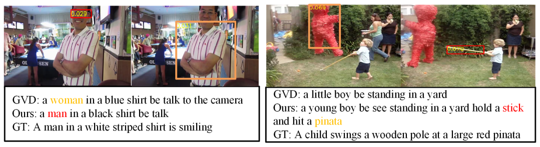
3. EXPERIMENTS

3.1. Experimental Setups

Dataset. We evaluate our proposed method on the ActivityNet Entities dataset, the standard benchmark for grounded video description. The ActivityNet Entities dataset contains about 52k video segments and 158k annotated bounding boxes. In terms of object classes of annotated bounding boxes, we adapt the setting of [5], with 432 object classes.

Metrics. Following the original evaluation protocol from [5], we use 5 standard language evaluation metrics, including Bleu@1 [16], Bleu@4 [16], METEOR [17] and CIDEr [18].

Implementation Details. Following [5], we use two CNNs [19, 7] to extract appearance and motion features of each frame for segments, respectively. The global video feature $\boldsymbol{v}_g$ of segments is obtained by using pooling operation on the appearance and motion features over frame dimension. We optimize the training with Adam [20] for 40 epochs and select the model with the best validation CIDEr score for testing. The batch size, learning rate and k are set to 80, 2e-3 and 10, respectively.



GVD: a woman in a blue shirt be talk to the camera
Ours: a man in a black shirt be talk
GT: A man in a white striped shirt is smiling

GVD: a little boy be standing in a yard
Ours: a young boy be see standing in a yard hold a stick and hit a pinata
GT: A child swings a wooden pole at a large red pinata

Fig. 3. Visualization of comparison between our method and the baseline. The solid line box and the number in the solid line box represent the most attended proposal region and the attention weight for the generated nouns in color and bold, respectively.

3.2. Ablation Studies

To verify the effectiveness of the proposed attention groundtruth sampling method and grounding supervision. We conduct ablation studies on the ActivityNet Entities validation set. Tab. 1 demonstrates the comparison of ablation experiments on the AE validation set. The results illustrate the boost brought by the proposed L_{attn} and IoP based sampling method. Note that the baseline GVD is denoted as $\hat{L}_{attn}$+L_{cls}+IoU. Comparing to the baseline, our method denoted by L_{attn}+L_{cls}+IoU+IOP gains an absolute gain of 1.20% in terms of CIDEr on the AE validation set. The baseline with IoP and the baseline with L_{attn} both do not gain improvement on grounded video description. It indicates small regions are essential for video description and the proposed grounding supervision is necessary to force the model to attend the discriminative part of the positive regions.

To investigate the impact of hyper-parameters. As show in Tab. 2 and Tab. 3, our method achieves the best performance when λ_{α}, τ_1 and τ_2 are set to 0.1, 0.5 and 0.9, respectively. In the following experiments, we maintain the above hyper-parameter setting if not otherwise specified.

3.3. Comparison with the State-of-the-arts

To demonstrate the effectiveness of our informative attention supervision method, we have conducted experiment by using our informative attention supervision method on a SOTA grounded video description models including GVD [5] and KNN-HAST [22].Tab. 4 shows the comparison of our method against existing approaches [7, 5, 21, 22, 23] on the AE dataset, our method consistently gains the improvement on several baseline models including GVD and KNN-HAST, which proves the compatibility of our method for the existing grounded video description models. Specifically, our method with KNN-HAST backbone provides a relative gain of more than 1.82% in terms of CIDEr and 3.57% in terms of Bleu@4 on the validation set of AE. Additionally, Our method provides a relative gain of more than 2.06% in terms of CIDEr compared to all previous works on the test set of AE. We believe that the improvement shown on multiple SOTA grounded video description models proves the effectiveness of our method.

Table 2. Comparisons of different settings of λ_{α} on the ActivityNet Entities validation set.

λ_{α}	λ_{β}	B@1	B@4	M	C
0.05	0.1	24.0	**2.83**	11.2	48.4
0.1	0.1	**24.0**	2.78	**11.2**	**48.7**
0.2	0.1	23.2	2.46	10.9	46.3
0.3	0.1	23.9	2.72	11.1	48.1
0.4	0.1	23.5	2.43	11.0	46.8
0.5	0.1	23.6	2.42	10.9	45.9

Table 3. Comparisons of different settings of τ_1 and τ_2 on the ActivityNet Entities validation set.

τ_1	τ_2	B@1	B@4	M	C
0.5	-	23.7	2.59	11.1	47.0
-	0.7	23.7	2.59	11.1	47.7
-	0.8	23.7	2.64	11.1	47.7
-	0.9	23.7	2.44	11.0	47.0
0.5	0.7	23.8	2.54	11.1	47.5
0.5	0.8	23.8	2.64	11.1	48.5
0.5	0.9	**24.0**	**2.78**	**11.2**	**48.7**

Table 4. Comparisons of the state-of-the-art methods on the ActivityNet Entities dataset.

	B@1	B@4	M	C
Validation set				
GVD [5]	23.9	2.59	11.2	47.5
Cyclical [21]	23.7	2.45	11.1	46.6
GVD(w/o SelfAttn) [5]	23.2	2.28	10.9	45.6
KNN-HAST [22]	-	2.80	11.3	49.4
Our Method(GVD)	24.0	2.78	11.2	48.7
Our Method(KNN-HAST)	**24.4**	**2.90**	**11.3**	**50.3**
Test set				
Masked Transformer [7]	22.9	2.41	10.6	46.1
Bi-LSTM+TempoAttn [7]	22.8	2.17	10.2	42.2
Cyclical [21]	23.4	2.43	10.8	46.6
GVD(w/o SelfAttn) [5]	23.1	2.16	10.8	44.9
GVD [5]	23.6	2.35	11.0	45.5
RGL [23]	**25.5**	2.59	11.0	47.4
KNN-HAST [22]	-	2.61	11.3	48.5
Our Method(GVD)	23.7	2.46	11.0	47.3
Our Method(KNN-HAST)	24.2	**2.76**	**11.3**	**49.5**

3.4. Qualitative Results

As shown in Fig. 3, compared to GVD, our method with GVD devotes more attention to the head, which is crucial for gender description and smaller than the body. So our method generates the correct word 'man' while GVD fails. Furthermore, as illustrated in the right terms of Fig. 3, the generated sentences our method is more precise because it contains more nouns that are small objects in frames.

4. CONCLUSION

In this paper, we propose an informative attention supervision method to improve the performance of GVD. The informative attention supervision method contains the proposal based attention groundtruth sampling method for providing more small regions that were previously neglected and the group based grounding supervision that guides the video description model to focus on the informative part of the sampled regions. Therefore, our method can boost the performance of existing video description models and provides better interpretability of the description sentences. Finally, experimental results demonstrate that our approach achieves a competitive description performance on the ActivityNet Entities dataset and reveal the small but informative regions are crucial for grounded video description.

5. REFERENCES

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[2] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 706–715.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[5] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach, "Grounded video description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6578–6587.

[6] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille, "Attention correctness in neural image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 4176–4182.

[7] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[8] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko, "Object hallucination in image captioning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4035–4045.

[9] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 771–787.

[10] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Neural baby talk," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7219–7228.

[11] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 563–579.

[12] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[15] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[17] Michael Denkowski and Alon Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Workshop on Statistical Machine Translation*, 2014, pp. 376–380.

[18] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[19] Yuanjun Xiong, Limin Wang, Zhe Wang, Bowen Zhang, Hang Song, Wei Li, Dahua Lin, Yu Qiao, Luc Van Gool, and Xiaoou Tang, "Cuhk & ethz & siat submission to activitynet challenge 2016," *arXiv preprint arXiv:1608.00797*, 2016.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira, "Learning to generate grounded visual captions without localization supervision," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 353–370.

[22] Kai Shen, Lingfei Wu, Fangli Xu, Siliang Tang, Jun Xiao, and Yueting Zhuang, "Hierarchical attention based spatial-temporal graph-to-sequence learning for grounded video description," in *International Joint Conferences on Artificial Intelligence*, 2020, pp. 941–947.

[23] Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang, "Relational graph learning for grounded video description generation," in *ACM International Conference on Multimedia*, 2020, pp. 3807–3828.