Visual Relation-Aware Unsupervised Video Captioning

Puzhao Ji[1], Meng Cao[1], and Yuexian Zou[1,2(✉)]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
zouyx@pku.edu.cn
[2] Peng Cheng Laboratory, Shenzhen, China

Abstract. Unsupervised video captioning aims to describe videos from unlabeled videos and sentence corpus without the reliance on human annotated video-sentence pairs. A straightforward manner is to borrow the merit from unsupervised image captioning methods, which resort to pseudo captions retrieved by visual concepts detected in image. However, directly applying this methodology to the video domain leads to sub-optimum performance since visual concepts cannot represent the major video content accurately and completely. Besides, these methods also do not consider the problem of *noise interference* caused by words unrelated to visual concept in the pseudo captions. In this paper, we propose a visual relation-aware unsupervised video captioning method which retrieves pseudo captions using visual relation. Based on these, we train the proposed visual relation-aware captioning model. Specifically, our model is designed to focus on learning from *dependable* words corresponding to the detected relation triplets. Extensive experimental results on two public benchmarks show the effectiveness and significance of our method.

Keywords: Video captioning · Visual relation · Unsupervised learning

1 Introduction

Video captioning seeks to automatically generate a sentence that describes the relation and interaction of objects in video. In applications, the video captioning model can describe the changes of people and objects around the visually impaired [1], foster and facilitate physical activities [2], automatically generate news releases for news videos [3], etc.

Most of the existing models [4–7] are trained in a supervised learning manner with human annotated video-sentence pairs (cf. Fig. 1(a)). However, manual annotation is very expensive and time-consuming due to the complex spatial and temporal dynamics of video. In addition, most of the annotations obtained by crowdsourcing are short and repetitive [8,10–13]. As a result, captioning models trained on crowdsourced annotation data have poor generalization in the wild. Therefore, it becomes increasingly important for video captioning models to get

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13531, pp. 495–507, 2022.
https://doi.org/10.1007/978-3-031-15934-3_41

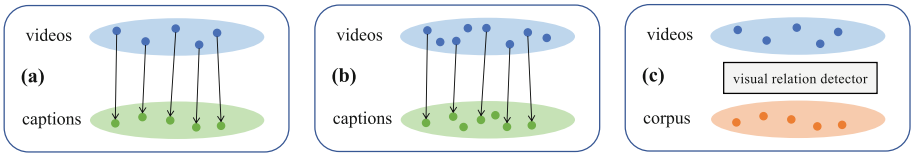

Fig. 1. Difference between visual captioning methods: (a) supervised video captioning [4], (b) semi-supervised video captioning [14,15], (c) our unsupervised video captioning.

rid of the annotated data. Recently, some semi-supervised approaches [9,14,15] have attempted to reduce reliance on annotated data. These methods use additional unlabeled video or sentences to train video captioning models. However, none of them gets rid of the dependency on human-labeled data (Fig. 1(b)). Therefore, in this paper, we address unsupervised visual captioning which only requires unlabeled data for training (Fig. 1(c)), making a more scalable solution under the large-scale easily accessible data.

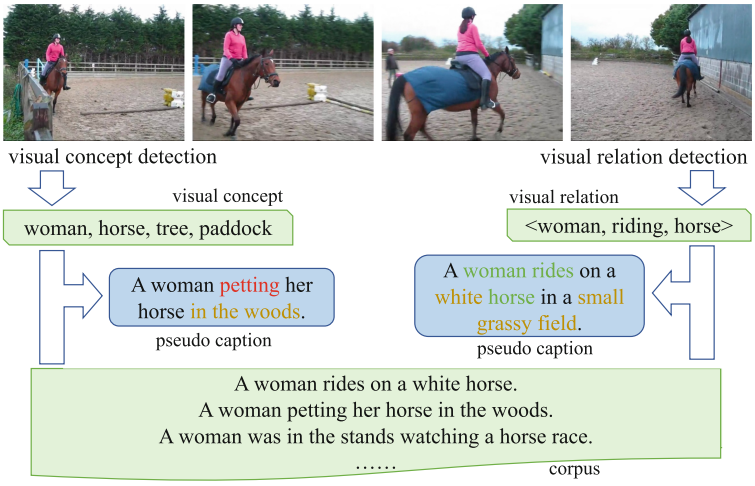

Fig. 2. The difference between the visual concept-based pseudo caption retrieve method and our method. Words in red, yellow and green indicate wrong relation between objects, noisy words and dependable words (words in visual relation triplet) respectively. (Color figure online)

It is noted that there are some works [8,10,16] are devoted to address unsupervised *image* captioning. These methods use the visual concepts detected in the images to match pseudo-labels from the corpus, and then train the decoders in the image captioning model using the pseudo-labels. However, we content that directly applies their visual concepts based methodologies to the video domain

leads to inferior performance. We declare this to the following two potential reasons. **1)** The pseudo-label retrieval method based on visual concept cannot capture the action and position relation of objects in the video. These visual concepts are actually detected objects. As shown in Fig. 2, such methods do not take into account the interaction between objects and they will introduce many irrelevant and even wrong pseudo-labels. **2)** The pseudo-labels retrieved have noise at the word level. In addition to the matched visual concepts, there are still a large number of words unrelated to the video content in the pseudo captions (such as 'in the woods', 'white', 'small grassy field' in Fig. 2).

We solve the above problems through the following two aspects: **1)** Use visual relation to retrieve pseudo-labels. Actually, a video mainly describes a major action or scene. Visual relation triplet ⟨`subject, relation, object`⟩ can properly express the main action or scene in the video. The `relation` refers to interaction or positional relation between objects. The pseudo-labels obtained by our method can be consistent with the main content of the video at the sentence level. **2)** To alleviate word level noise, we proposed Visual Relation-Aware Module (VRAM). VRAM makes the model learn from dependable words by giving higher confidence scores to the words in the visual relation triplet (such as green words: 'woman', 'rides', 'horse' in Fig. 2). This reduces the impact of noise at the word level.

The contributions of this paper are three-fold: **1)** We develop a visual relation-based pseudo caption retrieve mechanism which builds a bridge between video and pseudo captions. **2)** A module is devised to alleviate noise from words that are not part of matched relation triplet in pseudo caption. **3)** Extensive results on MSVD, MSR-VTT, and multiple corpus demonstrate the effectiveness of our method. Our visualization and results also show the domain and volume of corpus have a significant impact on the quality of retrieved pseudo captions.

2 Related

2.1 Supervised Video Captioning

The early works of video captioning extract fixed content like verb, subject and object, then populate the content into predefined template [17]. Withing fixed predefined template and limited hand-crafted grammar rules, these methods are hard to generate flexible and accurate description. Benefit from the raising of deep neural networks, sequence learning based methods [4–6] which adopt encoder-decoder framework, are widely used to describe video content with flexibility. Venugopalan et al. [4] propose a stacked LSTM model and average the feature of each video frame. Yao et al. [5] introduce a soft attention mechanism to capture the feature of salient frame or region. Chen et al. [6] proposes PickNet to choose key frames to reduce redundant visual information. More recently, RecNet [18] uses a reconstructor architecture to leverages the backflow from sentence to video while generating caption. Zheng et al. [7] introduce a SAAT module to generate syntax parts in caption. To employ the POS syntactic information, Wang et al. [19] propose a POS generator and use gating block to fuse multimodal feature. However, all the above methods rely one video-sentence annotation pairs.

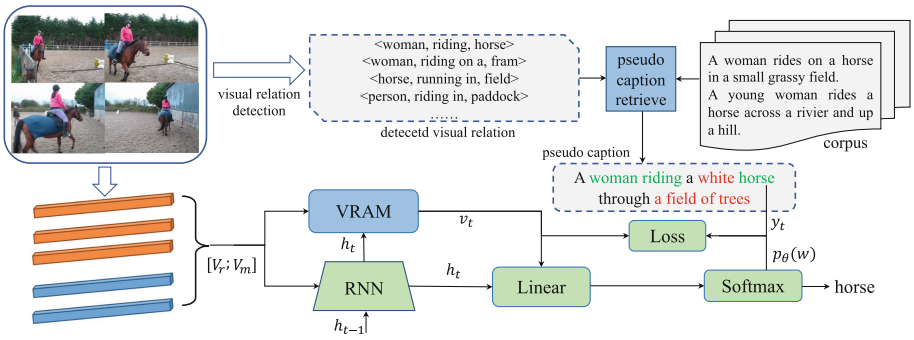

Fig. 3. The overview of our method, consisting of a pseudo caption retrieve mechanism and a visual relation-aware captioning model. Pseudo captions are matched by visual relation. The Visual Relation-Aware Module focus on dependable words that corresponding to visual relation. Words in green and red mean dependable words in relation triplet and video-irrelevant words. (Color figure online)

2.2 Unsupervised Captioning

Existing unsupervised visual captioning methods focus on image captioning, none of them are specifically designed to solve unsupervised video captioning task. Feng et al. [10] develop an architecture to align visual and textual features in common latent space to reconstruct each other, and train captioning model in adversarial manner using pseudo captions which based on visual concept. Laina et al. [8] project video and pseudo caption into a shared latent space structured by visual concepts, then decode caption from this latent space. However, these visual concept-based pseudo retrieve method fails to capture the relation between objects in video.

3 Method

An overall pipeline of proposed method is shown in Fig. 3. There are two components in our approach, a pseudo caption retrieve mechanism and visual relation-aware captioning model which consists of a language decoder and Visual Relation-Aware Module (VRAM). The captioning model learns from the results obtained by pseudo captions retrieve mechanism.

3.1 Pseudo Captions Retrieve

Compared with image which is still frame and contains limited objects, there may be environment changes and much more objects in video. In this case, visual concepts detected by the object detector are messy and cannot accurately represent the content in the video. However, the content of a video is a specific event that can be expressed as a visual relation triplet $t = \langle \mathtt{s,r,o} \rangle$ such as $\langle$ `woman,riding,horse` $\rangle$. s means the subject of the relation in triplet, r is the

relation, and o is the object of the relation. Similarly, the content of a sentence can also be condensed to relationships between objects.

We extract N frames from video v uniformly. For each frame, we use 2DCNN [29] and object detector [33] to get RGB feature V_r and ROI feature V_o. A pretrained visual relation detector [31] takes V_r and V_o as inputs and output the visual relation triplet t. We collect all visual relation triplet of N frames and get visual relation set $\mathcal{R}_v$ for video v. We construct a relation set $\mathcal{R}_y$ for each sentence y in corpus $\mathcal{C}$ by parsing the semantic and part-of-speech information. And the pseudo captions of this video is $\mathcal{Y} = \{y|\mathcal{R}_y \cap \mathcal{R}_v \neq \emptyset, y \in \mathcal{C}\}$. This ensures that each pseudo label has at least one identical visual relation triplet with video.

3.2 Basic Video Captioning Model

In this section, we describe the basic video captioning model directly trained by pseudo caption. Given a video, the RGB feature $V_r \in \mathbb{R}^{d_r}$ from 2DCNN and motion feature $V_m \in \mathbb{R}^{d_m}$ from 3DCNN are encoded into a single feature $V \in \mathbb{R}^{d_k}$ which is the input of basic video captioning model. The basic model directly take noisy pseudo caption $y = \{y_1, y_2, ..., y_n\}$ as the label. At time step t, the basic model generates word as follows:

$$V = \text{ReLU}(W_v[V_r; V_m]), \tag{1}$$

$$h_t = \text{RNN}(V, h_{t-1}), \tag{2}$$

$$p_\theta(w|h_t) = \text{softmax}(W_s \text{ReLU}(h_t)), \tag{3}$$

$$\hat{y}_t = \underset{w \in vocab}{\arg\max}\, p_\theta(w|h_t), \tag{4}$$

where $h_t \in \mathbb{R}^{d_h}$ is hidden state of RNN at time step t, $p(w|h_t)$ is generated word probability distribution, $W_s \in \mathbb{R}^{d_h}$ and $W_v \in \mathbb{R}^{d_k \times (d_v + d_m)}$ are learnable parameters, and the output word of basic model is $\hat{y}_t$. The parameter θ of basic model is optimized by the Cross-Entropy loss:

$$\mathcal{L}_c = -\frac{1}{n} \sum_{t=1}^{n} \log p_\theta(\hat{y}_t = y_t|h_{t-1}, V) \tag{5}$$

3.3 Visual Relation-Aware Module

This basic model works well with human annotations, but pseudo caption may contains words which are irrelevant to video or have ambiguous meaning. To alleviate the suboptimal training process by inaccurate label, we propose a Visual Relation-Aware Module (VRAM). VRAM designed to evaluate whether the hidden state of RNN at current time represents a dependable word or not. More specifically, VRAM estimate the probability that current word is part of a visual relation of input video. At time step t, a transformation matrix $W_p \in \mathbb{R}^{d_h \times d_k}$ used to project video feature to a feature space with dimension same as hidden

state and get the confidence score v_t of current hidden state. The confidence score of current hidden state generated as follow:

$$v_t = \text{sigmoid}(\frac{h_t^T W_p V}{\sqrt{d_h}}) \tag{6}$$

We limit v_t between 0 and 1 using sigmoid. When a hidden state gets a high confidence score, it will play a more important role in generating words. Using VRAM, the probability distribution of the generated word is as follows:

$$p_\theta(w|h_t, V) = \text{softmax}(W_s \text{ReLU}([v_t h_t; (1 - v_t)V])) \tag{7}$$

$$\mathcal{L}_v = -\frac{1}{n}\sum_{t=1}^{n}(v_t - v_t^*)^2 \tag{8}$$

$$v_t^* = \begin{cases} 1, & y_t \in t, t \in \mathcal{R}_v \cap \mathcal{R}_y \\ 0, & otherwise \end{cases} \tag{9}$$

The VRAM optimized by relation loss $\mathcal{L}_v$ in which v_t^* is constructed from pseudo caption to train VRAM. In backpropagation, larger v_t causes the model parameters learn more from y_t in pseudo caption. Combining captioning loss $\mathcal{L}_c$ and $\mathcal{L}_v$, we get loss function of our model.

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_v \tag{10}$$

4 Experiments

4.1 Datasets and Metrics

We take the videos in two widely used video captioning datasets: Microsoft Video Description Corpus (MSVD) [21] and Microsoft Research Video To Text (MSR-VTT) [20], and take corpus from the training split of TV show Caption (TVC) [22], Google's Conceptual Captions (GCC) [23], and VATEX [24]. VATEX and TVC are datasets of video captioning, but TVC is made to describe the tv shows, and GCC is a large image captioning dataset. We conduct experiments on them to compare the impact of corpus's domain on unsupervised video description.

Following the existing works [7], we split MSR-VTT as 6513 for training, 497 for validation, and 2,990 for testing. According to the common splits, MSVD is divided into 1200 for training, 100 for validation, and 670 for testing. The training split of TVC, GCC, and VATEX contain 182,556, 2,402,941 and 293,757 sentences respectively. Note that these descriptions do not overlap with the annotations of MSVD and MSR-VTT. We evaluate our method on the validation/test split of MSVD and MSR-VTT using widely-used metrics including BLEU@4 [25], METEOR [26], ROUGE_L [27] and CIDEr [28].

4.2 Implementation Details

For video feature extraction, we use ResNeXt [29] model and ECO [30] model which pretrained on ImageNet ILSVRC2012 dataset and Kinetics400 dataset respectively. By feeding 32 frames which sampled uniformly from video to ResNeXt and ECO, we get RGB feature of 2048 dimension and motion feature of 1536 dimension for each video.

For sentences in corpus, we remove unprintable characters, normalize punctuation, then apply part-of-speech tagging using spacy[1] to extract relation between objects. For extracted 32 frames of each video, we detect the visual relation in training split of MSVD and MSR-VTT using VDR-DSR [31], which results in 18,450 and 121,620 visual relation triplets. We use these visual relations to retrieve pseudo-labels from the corpus, taking into account the plural form of nouns and their synonyms. We adopt one layer GRU with hidden state of 512-dim as our language decoder. The word embeddings are initialized by random and embedding dimension is 512. We train our model with batch size of 32, and Adam [32] optimizer which initial learning rate is 0.001. We evaluate our model using beam search with size of 4 on test split.

4.3 Quantitative Results

We conducted extensive experiments on the MSVD dataset using three corpus. Following the common setting in unsupervised image captioning [16], the means and standard deviation of five runs results using random seed are represented. As shown in Table 1, the method "full model" is our full model, "w/o VRAM" represents GRU (without VRAM) trained by pseudo captions which retrieved by visual relations and "w/o relation" represents GRU trained by pseudo captions which retrieved by visual concepts. In "w/o relation" setting, the visual concepts are detected by FatserRCNN pretrained on OpenImage. For all three corpus, our method achieves the best results on CIDEr. It can be observed that the CIDEr's mean score of "w/o VRAM" is two times or more than the "w/o relation" on GCC and VATEX. This confirms the effectiveness of visual relation in pseudo captions matching. The results in Table 2 which is conducted on MSR-VTT using TVC corpus also support this observation. The comparison of "full model" with "w/o VRAM" demonstrates the ability of VRAM to alleviate the noise in pseudo label.

We observed that "w/o relation" achieved the best results on BLEU@4, METEOR and ROUGE_L on GCC dataset. We believe this is mainly due to there are a lot of sentences in GCC that have the same visual concepts as the video but they express completely different meanings. This resulted in the sentence generated by captioning model also have many same visual concepts as the video but have a completely different meaning with video. Concurrence-based metrics: BLEU@4, METEOR, and ROUGE_L are failed to measure their difference, while semantically based CIDEr can distinguish this situation.

[1] https://spacy.io.

Table 1. Performance comparison of different corpus and settings on the test split of MSVD dataset. The means and standard deviation of five runs results using random seed are represented.

Corpus	Method	BLEU@4	METEOR	ROUGE_L	CIDEr
TVC	full model	**5.3 ± 0.2**	**16.2 ± 1.1**	**40.6 ± 1.7**	**8.8 ± 0.9**
	w/o VRAM	3.6 ± 1.2	14.3 ± 1.1	39.9 ± 2.0	4.2 ± 0.8
	w/o Relation	4.2 ± 0.7	16.1 ± 0.6	40.0 ± 1.9	3.3 ± 1.8
GCC	full model	4.8 ± 0.8	14.2 ± 0.5	38.4 ± 1.8	**12.9 ± 2.0**
	w/o VRAM	1.4 ± 0.5	11.8 ± 1.1	32.5 ± 1.8	6.5 ± 2.4
	w/o relation	**5.6 ± 1.4**	**15.9 ± 0.7**	**40.7 ± 4.3**	2.8 ± 1.2
VATEX	full model	**9.6 ± 1.3**	**20.1 ± 0.9**	**42.1 ± 3.9**	**13.2 ± 0.8**
	w/o VRAM	6.7 ± 1.1	18.6 ± 0.7	39.2 ± 1.9	8.3 ± 0.9
	w/o relation	5.8 ± 1.7	16.5 ± 1.1	37.1 ± 4.9	1.3 ± 0.3

Table 2. Performance comparison on test split of MSR-VTT dataset using TVC corpus.

Method	BLEU@4	METEOR	ROUGE_L	CIDEr
full model	**8.4 ± 0.5**	**15.9 ± 0.7**	37.8 ± 1.0	**4.9 ± 0.4**
w/o VRAM	3.5 ± 1.1	15.2 ± 1.6	**42.6 ± 1.5**	2.4 ± 0.5
w/o relation	5.1 ± 1.8	12.6 ± 1.3	32.5 ± 2.8	1.1 ± 0.3

The results in Table 1 also indicate that the domain and volume of corpus have a significant impact on the quality of retrieved pseudo captions. TVC corpus is consists of captions that describe tv shows. And its content is quite different from MSVD which contains web videos of various categories. GCC is an image caption dataset, its captions tend to describe still frame. The results of GCC corpus works better than TVC on three settings in case that the two corpus do not match the domain of MSVD very well. This is because GCC has 13 times more sentences than TVC, which greatly increases the likelihood of appearance of video content related sentences. VATEX achieves the best performance in three corpus because it is also an open domain video captioning dataset.

Table 3. Ablation study on corpus which in same domain but has different volume.

Corpus	BLEU@4	METEOR	ROUGE_L	CIDEr
Oracle	27.4 ± 2.4	24.6 ± 0.7	60.4 ± 1.5	30.9 ± 2.6
MSR-VTT	13.5 ± 0.8	20.5 ± 1.2	47.5 ± 5.9	18.0 ± 1.7
VATEX	9.6 ± 1.3	20.1 ± 0.9	42.1 ± 3.9	13.2 ± 0.8
MSR-VTT+VATEX	**16.4 ± 4.7**	**21.6 ± 1.7**	**51.3 ± 7.0**	**18.5 ± 2.1**

To further investigate the effect of corpus's size, we conduct ablation experiments on MSVD with MSR-VTT, VATEX, and MSR-VTT+VATEX. It is noted that MSVD, MSR-VTT and VATEX are all open domain video captioning datasets. As shown in Table 3, the MSR-VTT+VATEX outperforms MSR-VTT and VATEX on all four metrics, which demonstrates that our method retrieves more video-related but diverse pseudo captions from a larger corpus. And following the common practice [8], we also use Oracle corpus to see if our method has the ability to retrieve better-quality pseudo captions from corpus which are more relevant to video content. The Oracle means retrieve pseudo labels from the ground truth of MSVD training set. The BLEU@4, METEOR, ROUGE_L and CIDEr of Oracle corpus both outperformed the other corpus which confirmed that correlation between corpus and videos had a great influence on the results of unsupervised video captioning. And Oracle corpus only used to show the upper limit of the ability of visual relation in pseudo label matching.

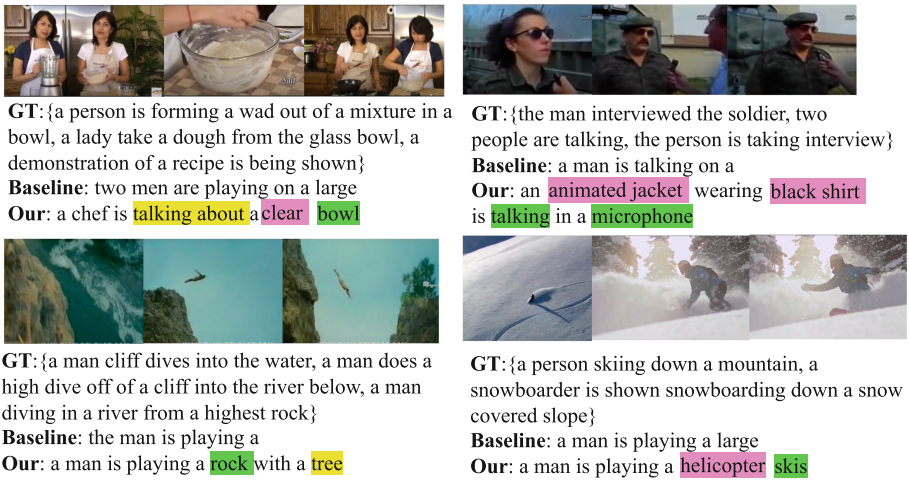

Fig. 4. The qualitative results between our method and visual concept based baseline. Words with red, yellow, and green backgrounds means complete irrelevant words, words that ambiguously used, and words that correctly used, respectively. (Color figure online)

4.4 Qualitative Results

Figure 4 shows four qualitative results generated by our method and visual concept-based method. In these examples, the baseline failed to capture the major content of the video, and even cannot generate correct sentence structure. Our method successfully output words such as "bowl", "talking in microphone" and "skis", but the noise in the pseudo label also cause words with unclear or wrong meanings such as "helicopter" and "animated jacket". There are also some ambiguous words generated, such as "talking about" in the upper-left example and "tree" in the bottom-left example. Limited by the detection range of visual

relation detector, our method fails to recognize the object "cliff" and action "dive" in bottom-left video.

Finally, to provide more insight about retrieved pseudo labels, we visualize the t-SNE embedding of one video's pseudo captions. As shown in Fig. 5, although the pseudo captions retrieved by visual relation are closer to the human annotated labels, there is still a large gap between them. It can be inferred that pseudo caption matched by visual relation lack of details about video content.

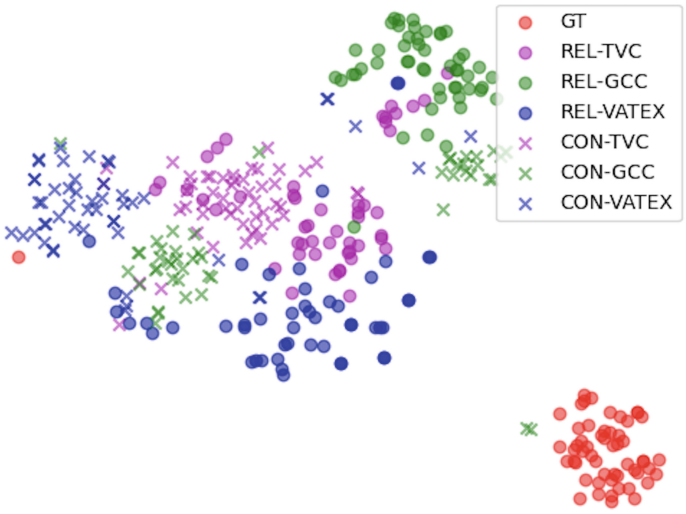

Fig. 5. t-SNE embedding of pseudo captions and ground truth. REL-X and CON-X denote pseudo captions retrieved from corpus X using visual relation and visual concept respectively. GT means human annotated captions.

5 Conclusion

In this paper, we make the first attempt to investigate unsupervised video captioning. For this purpose, we propose **1)** a new visual relation-based pseudo caption retrieve method to match the major content of video. **2)** Visual Relation-Aware Module(VRAM) to mitigate extra noisy caused by words irrelevant to visual relation. Our experiments show results beyond visual concepts-based method and we also show intuitive visualization of pseudo captions. In the future, we will further investigate the semantic structure of video and sentence, and design more effective pseudo label matching mechanism and module.

Acknowledgements. This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science and Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References

1. Huang, Y.H., Hsieh, Y.Z.: The assisted environment information for blind based on video captioning method. In: IEEE International Conference on Consumer Electronics, pp. 1–2 (2020)
2. Amirian, S., Farahani, A., Arabnia, H.R., Rasheed, K.M., Taha, T.R.: The use of video captioning for fostering physical activity. In: International Conference on Computational Science and Computational Intelligence, pp. 611–614, (2020)
3. Whitehead, S., Ji, H., Bansal, M., Chang, S.F., Voss, C.R.: Incorporating background knowledge into video description generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3992–4001 (2018)
4. Darrell, T., Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Saenko, K.: Sequence to sequence - video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)
5. Yao, L., et al.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
6. Chen, Y., Wang, S., Zhang, W., Huang, Q.: Less is more: picking informative frames for video captioning. In: Proceedings of the 15th European Conference on Computer Vision, pp. 367–384 (2018)
7. Wang, C., Zheng, Q., Tao, D.: Syntax-aware action targeting for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 13093–13102 (2022)
8. Laina, I., Rupprecht, C., Navab, N.: Towards unsupervised image captioning with shared multimodal embeddings. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7413–7423 (2019)
9. Zhang, C., Yang, T., Weng, J., Cao, M., Wang, J., Zou, Y.: Unsupervised pretraining for temporal action localization tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 14031–14041 (2021)
10. Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4120–4129 (2019)
11. Zhang, C., Cao, M., Yang, D., Chen, J., Zou, Y.: CoLA: weakly-supervised temporal action localization with snippet contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16010–16019 (2021)
12. Cao, M., Chen, L., Shou, M.Z., Zhang, C., Zou, Y.: On pursuit of designing multimodal transformer for video grounding. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 9810–9823 (2021)
13. Zhang, C., Yang, T., Weng, J., Cao, M., Wang, J., Zou, Y.: Deep motion prior for weakly-supervised temporal action localization. arXiv preprint arXiv:2108.05607 (2021)
14. Lucchi, A., Chen, W., Hofman, T.: A semi-supervised framework for image captioning. arXiv preprint arXiv:1611.05321 (2016)
15. Oh, T., Kim, D., Choi, J., Kweon, I.: Image captioning with very scarce supervised data: adversarial semisupervised learning approach. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2012–2023 (2019)

16. Hashimoto, A., Watanabe, T., Honda, U., Ushiku, Y., Matsumoto, Y.: Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, pp. 3692–3702 (2021)
17. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. Int. J. Comput. Vision **50**, 171–184 (2002)
18. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7622–7631 (2018)
19. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video captioning with POS sequence guidance based on gated fusion network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2641–2650 (2019)
20. Yao, T., Xu. J., Mei, T., Rui, Y.: MSR-VTT: a large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)
21. Malkarnenkar, G., et al.: Youtube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2712–2719 (2013)
22. Bansal, M., Lei, J., Yu, L., Berg, T.L.: TVQA: localized, compositional video question answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1369–1379 (2018)
23. Goodman, S., Sharma, P., Ding, N., Soricut, R.: Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2556–2565 (2018)
24. Chen, J., Li, L., Wang, Y., Wang, X.E., Wu, J., Wang, W.Y.: VATEX: a large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4580–4590 (2019)
25. Ward, T., Papineni, K., Roukos, S., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
26. Denkowski, M.J., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
27. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: The Association for Computational Linguistics (ACL) Workshop, vol. 8 (2004)
28. Zitnick, C.L., Vedantam, R., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)
29. Dollar, P., Tu, Z., Xie, S., Girshick, R.B., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5987–5995 (2017)
30. Singh, K., Zolfaghari, M., Brox, T.: ECO: efficient convolutional network for online video understanding. In: Proceedings of the 15th European Conference on Computer Vision, pp. 713–730 (2018)
31. Chang, H., Liang, K., Guo, Y., Chen, X.: Visual relationship detection with deep structural ranking. In: Proceedings of the Thirty-Second Conference on Artificial Intelligence, pp. 7098–7105 (2018)

32. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (2014)
33. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2015)