Journal of Advanced Marine Engineering and Technology, Vol. 47, No. 6, pp. 405~410, 2023
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2023.47.6.405

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Target-centered context-detection technique using dual R-CNN

Jin-Se Lee[1] · Seong-Beom Jeong[2] · Yeong-Jae Shin[3] · Dong-Hoan Seo[†]

(Received December 8, 2023 ; Revised December 16, 2023 ; Accepted December 21, 2023)

Abstract: Image captioning, which aims to understand the context of an image, generates natural language using object feature vectors. This results in verbose sentences because they usually contain modifiers and objects. However, unlike human descriptions, the information needed in natural environments must be simplified. Therefore, we propose a target-centered context-detection model that uses dual R-CNN to generate short sentences about subject-centered attributes to describe the behavior of the target. The proposed model consists of target context detection (TCD), which detects subjects and actions, and activity image caption, which generates sentences centered on actions. The proposed TCD uses two RCNN heads to estimate objects and their properties. In this process, we added target-description region expansion to filter out unnecessary objects and encompass surrounding information. Afterward, the proposed activity image caption combines the feature vector and attributes of each target to generate a short description of the target-attribute pair. The proposed model can effectively convey information using brief rather than long sentences.

Keywords: Dense image captioning, Faster R-CNN, Target-description region expansion

1. Introduction

CCTV has become an essential element of modern urban centers used to prevent crime and maintain public order in public places. Existing surveillance systems only serve as records for follow-up management due to limitations in operating personnel. Recently, several studies have proposed technologies such as object detection and tracking to make these systems completely unmanned. However, these systems still play a limited role because they have no judgment function. This surveillance system requires an understanding of objects for proactive response.

Cameras can acquire vast amounts of information compared with other sensors. Many systems take advantage of this capability and apply it to various tasks. Some studies suggest methods of processing visual information that combine it with language models to process large amounts of information. Image captioning is creating a single sentence describing the primary situation of the image. Vinyals *et al.* **[1]** showed that images can be used as language based on deep learning. Since then, various studies have focused on generating sentences with more accurate and diverse expressions. However, when humans read, they prefer concise and accurate sentences that contain diverse information rather than rich expressions. This characteristic is essential for surveillance systems and industrial solutions such as CCTVs.

Johnson *et al.* **[2]** provided detailed information about each object by individually generating sentences from various objects in the image. They proposed a DenseCap with a structure that combines an object-detection network and an image-captioning network. The dense image captioning had the same encoder-decoder structure as existing image captioning. Because dense image captioning requires various object information from one image, captioning for each object is possible by extracting objects from the image as feature vectors at the encoder stage and passing them to the decoder. However, in this process, only a minimal area and associated objects, including surrounding information, are removed, making it challenging to obtain behavioral information about the object.

† Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics & Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412

1 M. S., Department of Electrical and Electronics Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: ljs171181@g.kmou.ac.kr, Tel: 051-410-4822

2 M. S., Department of Electrical and Electronics Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: sincere96@g.kmou.ac.kr, Tel: 051-410-4822

3 M. S., Department of Electrical and Electronics Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: yjshin0329@gmail.com, Tel: 051-410-4822

Yang *et al.* **[3]** and Li *et al.* **[4]** analyzed entire images and acquired the background and context information of the objects that were lost while extracting them as feature vectors. This method can depict the interaction between an object and the surrounding background but has the limitation of outputting only local information. Shin *et al.* **[5]** proposed a description region expansion (DRE) that creates a cluster object by integrating highly related neighboring objects based on intersection over union (IoU) before passing the feature vector to the decoder. In this method, clustered objects are considered images and the interactions between each object and surrounding objects are described in sentences.

Sentences generated through dense image captioning are generally verbose because they include modifiers for the object and surrounding information. However, the information needed in natural environments, including surveillance systems, is more about objects and their actions rather than lengthy descriptions. In particular, a description of a specific target is necessary to quickly understand the information displayed on the screen.

To solve this problem, we propose a target-centered context-detection technique using dual R-CNN, which aims to generate short sentences about attributes centered on the subject rather than the focus of the existing encoder-decoder structure. The proposed model connects two Faster R-CNNs before and after the target DRE (T-DRE), integrates them through object detection, target classification, and IoU comparison processes, and detects attributes with the integrated results. In this process, unnecessary objects are filtered out. The proposed target-context detection (TCD) can effectively convey information by expressing target objects and actions in the image.

2. Related Works

2.1 Faster R-CNN

Faster R-CNN **[6]** is a widely used object-recognition model from the R-CNN family. This model is the first end-to-end model in the R-CNN family. Due to its intuitive architecture, separated into region proposal network (RPN) and Fast R-CNN, object tracking **[7]-[9]**, image captioning **[10]-[12]**, OCR **[13]-[15]**, and so on, are widely used in many image fields. RPN extracts features using a single convolution layer for feature vectors obtained through the backbone network and uses a regressor layer to detect the object location and a classification layer to detect the probability that the object exists. The area where the object will exist and the probability that the object will exist in that area are detected. area compares The degrees of overlap between each detected area with the previously created anchor box are compared. The final area where the object is located, and the classification information of the object existing within that area are extracted by Fast R-CNN.

2.2 Description Region Expansion

Existing dense image captioning acquires object information from object recognition to describe multiple objects within one image. Here, the bounding box of the object has only a minimal area. As a result, the sentence lacks surrounding information and only describes information about local objects. To solve this, Shin *et al.* **[5]** analyzed the correlation between adjacent objects by integrating the bounding boxes of objects based on IoU.

The DRE analysis method first sorts bounding boxes in order of highest objectness score. Afterward, the size of the bounding box is expanded by 10%, and the IoU of the highest bounding box and the remaining bounding boxes are calculated. Here, bounding boxes with IoUs of 0.05 or higher are integrated to create new bounding boxes. This process proceeds for all detected bounding boxes. Through this process, the newly created bounding boxes become cluster objects that include the surrounding background, allowing object characteristics to be clearly described.

3. Proposed Technique

3.1 Overview of the Proposed Technique

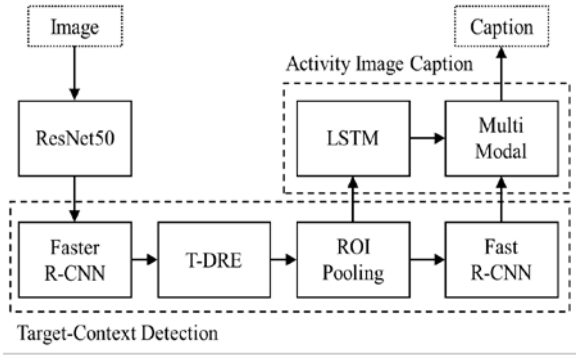

Figure 1: Structure of proposed target-centered context-detection technique

Existing image captioning may provide unnecessary information because it generates sentences by analyzing objects in the entire image. To solve this problem, we propose a TDC technique that describes the target and its actions by integrating surrounding objects with the target. TCD consists of two Faster R-CNNs

that extract objects and actions and T-DRE that integrates the extracted objects. Faster R-CNN can only extract objects from a minimal object area. As a result, the objects lose surrounding information, and behavioral information cannot be extracted. T-DRE creates cluster objects by integrating surrounding objects into the target object, among the objects extracted by R-CNN. Faster R-CNN further extracts the behavior of the target object from the cluster object. **Figure 1** shows the overall architecture of the proposed technique.

Images are input into the network with ResNet50 as the backbone and converted into feature vectors. Afterward, the object is detected by Faster R-CNN, and surrounding objects are integrated by T-DRE to create a group of objects targeting humans. The bounding box of the object now includes the surrounding area and is further expanded to include surrounding information. The object interacts with other objects and the surrounding environment. Therefore, we can estimate more accurate behavior by including surrounding information. Afterward, accurate target-context information can be extracted by pooling this bounding box to extract local feature information and re-estimate it based on the behavior and activity captions using Faster R-CNN, the head of R-CNN, and long short-term memory (LSTM).

3.2 Target-Description Region Expansion

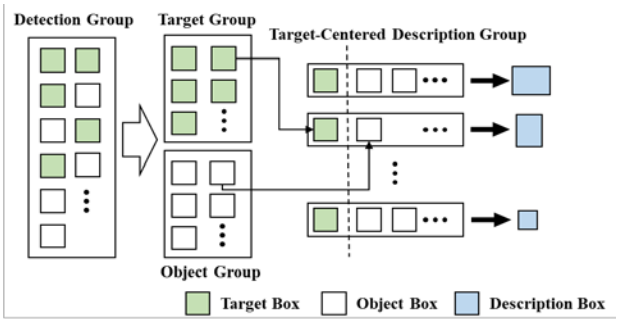

Figure 2: Flow of target-description region expansion

To describe the behavior of objects in an image, all objects in the image must first be recognized. For this purpose, Faster R-CNN, which has the advantage of high object-detection accuracy, is used to predict the type and location of objects in the image. The detected objects have a minimum region of interest, limiting their behavior description. To describe their behavior, we need to analyze the surrounding information and interactions. For this purpose, the proposed T-DRE integrates surrounding objects with the target object, and calculates the IoU between the target object and surrounding objects, judges it as information surrounding the target, and integrates it with the target object. **Figure 2** shows the schematic of the T-DRE process.

First, objects extracted from images by Faster R-CNN are the upper left coordinates of the bounding box, and w and h are the width and height of the box, respectively. The detected objects are sorted based on the prediction score, and objects below a specific value are discarded. It is possible to rule out false positives that occur during the object-detection process; therefore, the target and objects can be clearly distinguished. Detected objects are divided into target, the subject of the action, and object, which assists in explaining the action with surrounding information. Separated targets and objects are sorted into Target and Object groups. Afterward, the degree of overlap between the two boxes is analyzed using IoU calculation and the object box belongs to the object group, focusing on the target box. The IoU calculation is the union of the two bounding boxes divided by the intersection and has a value of between 0 and 1. The closer it is to 1, the greater the degree of overlap. IoU is calculated using the subject box and object box, and a target-centered description group (T-CDG) is created by collecting objects with an IoU of 0.7 or higher to ensure that objects and subjects are not too far apart. T-CDG creates a new bounding box, the description box, with the minimum value at the top left and the maximum value at the bottom right of the coordinates of each bounding box.

3.3 Activity Image Caption Based on Target Object

The size and ratio of the description box output from T-DRE are matched to the output size of Faster R-CNN. Therefore, to extract features from only the ratio of the description box in the existing feature map, the coordinates of the description box must be fixed on the feature map.

The feature map uses the output of Resnet50 FPN, the backbone of Faster R-CNN used as a detection network. The extracted feature map had dimensions of 256 × 13 × 13, and since the description box needed to be adjusted to this, a 13 × 13 grid that fitted the entire image was created to calculate the box ratio. The calculated size of the box crops the feature map, and this map is finally floated to interpret the description. For this purpose, the size of the cropped feature map was transformed to 256 × 7 × 7.

3.4 Language Model

The first feature map is transformed into a sentence in the form of the first embedding vector through the embedding layer and is sequentially input through the LSTM structure. The overall structure of the embedding layer is the same as **Equations (1), (2)** and **(3)**.

Table 1: Metric results comparison between the proposed model and others

Model	BLEU1	BLEU2	BLEU3	BLEU4	Meteor
FCLN **[2]**	-	-	-	-	0.31
CAG-NET **[19]**	-	-	-	-	0.32
ASG **[20]**	-	-	-	0.18	0.21
DRE **[5]**	**0.68**	0.447	0.252	0.14	0.46
Proposed Model	0.65	**0.50**	**0.39**	**0.32**	**0.73**

$$x_{t-1} = feature\ vector \tag{1}$$

$$x_t = S_t, t\epsilon\{0, \cdots, N-1\} \tag{2}$$

$$y_{t+1} = LSTM(x_t), t\epsilon\{0, \cdots, N-1\} \tag{3}$$

Here, S_0 is the start token indicating the input of the sentence, and S_N is the end token indicating the end of the sentence. The output of Faster R-CNN with behavioral information is then added as the input to the embedding layer of the LSTM. In this way, the output of the LSTM is finally connected to the fully connected layer to select the highest predicted value among the 1,791 indexed tokens and connect the sentences.

Table 2: mAP results between the proposed model and others

Model	mAP (%)
FCLN	5.39
CAG-NET	10.51
COCG **[4]**	8.90
Proposed Model	**36.85**

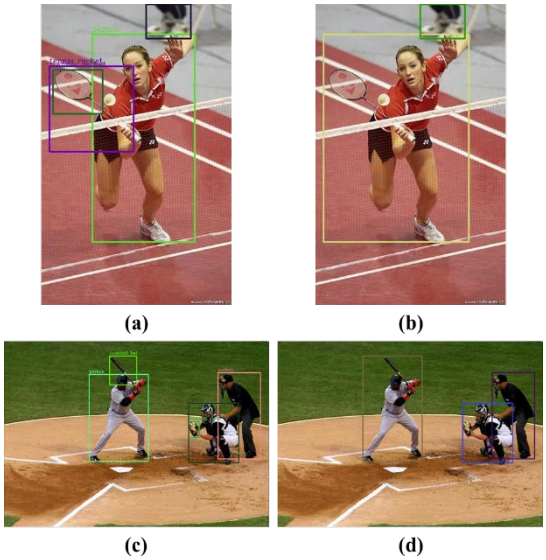

Figure 3: Result of target-description region expansion

4. Experiment and Result

4.1 Dataset and Preprocessing

This study used part of the Visual Genome (VG) dataset **[16]** to learn and verify a description model targeting humans. Among the subjects in the VG dataset, nine types of subjects related to people, including 'man,' 'person,' and 'people', were selected as targets. The VG dataset used contained 39,000 images depicting the target and 260,000 related sentences, and each sentence had relational properties such as target behavior and interaction with the object. We optimized the dataset by replacing words with an occurrence frequency of five or less with tokens during the dataset preprocessing. In addition, the start and end tokens, <bos> and <eos>, respectively, were placed at the beginning and end of the sentence so that the beginning and end of the entire sentence could be known. Lastly, the dataset was unified by placing <pad> tokens to match the input data size by matching short and long sentences. The total number of words was 1,791, including <pad>, <bos>, <eos>, and <unk>, and was used as the dimension of the final output layer.

4.2 Experiment and Result

The proposed method narrows the area of interest to the center of the target, excludes unnecessary relationships between objects, and implements target-centered context detection that increases accuracy and reduces learning difficulty by explaining only the object of interest. To evaluate the performance of the proposed model, we used BLEU 1–4 **[17]**, which compares consecutive equivalence with reference sentences in the dataset, and Meteor **[18]**, which includes synonyms. The higher the BLEU and Meteor the better the results. **Table 1** shows the scores for BLEU 1–4 and Meteor of the proposed model, the previous DRE paper, and the three models being compared. In **Table 1**, the BLEU-1 indicator received a lower score than the existing model, but other scores were relatively higher. This occurred because learning was conducted using 1,791 tokens, only 25% of the 6,665 tokens used for learning in the DRE paper, and evaluation was conducted using a limited number of tokens. However, in the

Meteor score, the proposed model received the highest score of 0.73, significantly higher than the compared models.

Also, in contrast to general captioning, dense image captioning describes objects; therefore, we must consider not only the accuracy of object detection but also the accuracy of the generated sentences **[2]**. To consider object accuracy, IoU and Meteor together, we increased the threshold of IoU from 0.3 to 0.7 in steps of 0.1, and the threshold of Meteor from 0.05 to 0.25 in steps of 0.05 to calculate mean Average Precision (mAP). The higher the mAP, the better the results. **Table 2** compares the mAP of the proposed model and the other three models measured by the above methods. From **Table 2**, we can see that the mAP of the proposed model is at least three times higher than other models. The reason for the high mAP is that the T-DRE model merges objects around the target object. In addition, the center of the description region is limited to the target object, which affects the detection, resulting in a higher mAP.

Figure 3 shows the results of a description box that integrates objects centered on the target object. **Figures 3 (a)** and **(c)** show the detection group output using Faster R-CNN on the image, while **(b)** and **(d)** show the results of T-CDG. T-DRE can combine the racket, bat, and glove of the target object to create a new description box. In this way, T-DRE can generate a description box of the interaction of the target object with an object or multiple objects, allowing for accurate sentence generation.

5. Conclusion

This paper proposes a target-centered context-detection technique in image captioning that describes sentences centered on the actions of the target. Existing captioning is lengthy because it includes modifiers and objects. The proposed TCD detects targets in images through two R-CNNs and describes the behavior of the detected objects. To describe behavior, the behavioral information of the target can be extracted by integrating surrounding information with the target through T-DRE. As a result, it is possible to accurately obtain information from surveillance systems such as CCTV by describing the behavior of the target instead of the existing lengthy sentences.

Through this study, it will be possible to augment data on human behavior in the future to describe behavior accurately. This is expected to prevent accidents and respond quickly on behalf of workers manning the surveillance system.

Acknowledgements

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 21HCLP-C162922-01).

Author Contributions

Conceptualization, S. B. Jeong and Y. J Shin; Methodology, J. S. Lee; Software, J. S. Lee and Y. J. Shin; Data curation J. S. Lee and S. B. Jeong; Writing-Original Draft Preparation, J. S. Lee; Writing-Review & Editing, D. H. Seo; Supervision, D. H. Seo.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 3156-3164, 2015.

[2] J. Johnson, A. Karpathy, and L. Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565-4574, 2016.

[3] L. Yang, K. Tang, J. Yang, and L. J. Li, “Dense captioning with joint inference and visual context,” Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2193-2202, 2017.

[4] X. Li, S. Jiang, and J. Han, “Learning object context for dense captioning,” Proceedings of the AAAI conference on Artificial Intelligence, vol. 33, no. 01, pp. 8650-8657, 2019.

[5] Y. J. Shin, S. B. Jeong, J. H. Seong, and D. H. Seo, “Description region expansion-based relationship-oriented dense image captioning model,” Journal of Advanced Marine Engineering and Technology (JAMET), vol. 45, no. 6, pp. 434-441, 2021.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, vol. 28, pp. 91-99, 2015.

[7] Z. Li, L. Zhang, Y. Fang, J. Wang, H. Xu, B. Yin, and H. Lu, “Deep people counting with faster R-CNN and correlation tracking,” in Proceedings of the International Conference on Internet Multimedia Computing and Service, pp. 57-60, 2016.

[8] F. Cui, M. Ning, J. Shen, and X. Shu, “Automatic recognition and tracking of highway layer-interface using Faster R-CNN,” Journal of Applied Geophysics, vol. 196, p. 104477, 2022.

[9] M. Othmani, “A vehicle detection and tracking method for traffic video based on faster R-CNN,” Multimedia Tools and Applications, vol. 81, no. 20, pp. 28347-28365, 2022.

[10] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 684-699, 2018.

[11] X. Yang, Y. Liu, and X. Wang, “Reformer: The relational transformer for image captioning,” in Proceedings of the 30th ACM International Conference on Multimedia, pp. 5398-5406, 2022.

[12] J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp. 4467-4480, 2019.

[13] Y. Nagaoka, T. Miyazaki, Y. Sugaya, and S. Omachi, “Text detection by faster R-CNN with multiple region proposal networks,” in Proceedings of 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 6, pp. 15-20, 2017.

[14] J. Yang, P. Ren, and X. Kong, “Handwriting text recognition based on faster R-CNN,” in Proceedings of 2019 Chinese Automation Congress (CAC), pp. 2450-2454, 2019.

[15] N. P. Ap, T. Vigneshwaran, M. S. Arappradhan, and R. Madhanraj, “Automatic number plate detection in vehicles using faster R-CNN,” in Proceedings of 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1-6, 2020.

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, and *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” International Journal of Computer Vision, vol. 123, pp. 32-73, 2017.

[17] 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 of the Association for Computational Linguistics, pp. 311-318, 2002.

[18] 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, pp. 65-72, 2005.

[19] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, “Context and attribute grounded dense captioning,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6241-6250, 2019.

[20] S. Chen, Q. Jin, P. Wang, and Q. Wu, “Say as you wish: Fine-grained control of image caption generation with abstract scene graphs,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9959-9968, 2020.