Journal of Advanced Marine Engineering and Technology, Vol. 46, No. 6, pp. 422~429, 2022
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2022.46.6.422

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

Activity-oriented visual relationship detection technique for context recognition

Yeong-Jae Shin[1] · Seong-Beom Jeong[2] · Ju-Hyeon Seong[3] · Dong-Hoan Seo[†]

(Received December 21, 2022 ; Revised December 26, 2022 ; Accepted December 26, 2022)

Abstract: Context recognition is a technology that acquires information about events based on information on various objects appearing in images. To implement this, dense image captioning, which recognizes all objects in an image, is often applied. However, because this approach targets all objects, it provides status information, such as location or color, which is relatively less important to humans, and even static object information. To humans, this information is unnecessary because of its low readability. To solve this problem, we propose a Context Pair Network that describes only the context of an object based on visual relationship detection. The proposed model consists of a pair object module (POM) that extracts subjects, objects, and relationships and a pair embedding module (PEM) that creates a subject–predicate–object structure. The proposed POM detects objects corresponding to subjects and objects based on three RCNNs and matches the detected objects with subject–object (S-O) pairs. The PEM also generates sentences consisting of a subject–predicate–object using S-O pairs based on long short-term memory. Thus, the proposed model is capable of situational awareness that provides only interactions between objects, unlike conventional captioning, which describes all the information indiscriminately.

Keywords: Visual relationship detection, Object detection network, Activity, Region of interest

1. Introduction

Recently, heightened social interest in public order and daily safety has expanded into a security system for its management. Video equipment, such as CCTV, a traditional security medium, is still mainstream. However, because this surveillance system relies on human, a surveillance gap arises as the system increases in size. Therefore, research on unmanned surveillance systems is actively being conducted worldwide. Therefore, many artificial intelligent (AI) solutions offer object detection, tracking, and anomaly detection. However, there is a limit to the level of performance that is comparable to that of humans. Recently, image captioning technology that describes the situation of an image as a natural language sentence has been introduced.

Image captions, which understand the context of an image and express it in natural language, have recently attracted significant interest as a technology that fuses two different pieces of information. In particular, this technology is very challenging because it aims at sentences equivalent to that of humans. Early image captions had a simple form of constructing sentence templates and inserting appropriate words **[2][3][4]**. However, since the introduction of deep learning, image and language information can be combined and created without a separate template. Therefore, current image captioning can accurately and richly describe images based on the reasoning ability and massive datasets of deep learning.

The modern image caption model consists of an area that interprets images and an area that creates sentences similar to the characteristics mentioned above. This encoder–decoder model

† 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 & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: yjshin0329@gmail.com, Tel: 051-410-4822

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

3 Assistant Professor, Department of Liberal Education & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: jhseong@kmou.ac.kr, Tel: 051-410-5031

uses convolutional neural networks (CNNs) to extract visual information and recurrent neural networks (RNNs) as decoders to generate sentences based on language information. The neural image caption (NIC) model **[5]** was the first image captioning model to use this structure. This model uses the Inception V3 model as an encoder to extract visual feature vectors from images and a decoder designed with long short-term memory (LSTM) to generate sentences. Since then, image captioning has used robust object recognition networks as encoders or improved decoders using new model structures. Recently, transformers have been primarily used as decoders, and many studies have been published **[6]**.

Through this structure, image captioning can remove unnecessary information by providing information in sentences that is easy for humans to understand. Because the feature vector extracted by the encoder contains information about the entire image, the sentences generated based on it tend to either contain information about the entire image or are based on the most prominent object. However, it is difficult to accurately express the information in a video with image capturing, which is output in only one sentence, because the video includes many objects, and each piece of information is often different. To solve this problem, Johnson *et al*. **[1]** proposed a dense image capturing technology that generates sentences for various objects in images by combining image captioning and object detection models.

Dense image captioning creates descriptions for each object detected in an image, and considerable information can be obtained from one image. The fully convolutional localization network (FCLN) **[1]**, the first dense image captioning model, detects objects present in images through faster-RCNN, extracts feature vectors of the objects, and generates sentences for each. Thus, information on the plurality of objects detected in one image can be expressed as individual sentences. However, because the object recognition model included in the FCNL model extracts a region of interest suitable for the size of the detected object, it tends to output only local information regarding the target object as a sentence. In addition, dense image captioning provides local information such as object color and state.

This indiscriminate information significantly assists in areas such as Q&A using images. However, this information is only necessary for domains that target implied information such as surveillance systems. Information regarding an active object in which two or more objects interact is required in a natural environment. Generally, in natural language, these are referred to as subjects and objects. In addition, the mutual relationship corresponding to the predicate can be defined in various ways. Visual relationship detection (VRD) classifies objects in an image into subject and object and derives a relationship between the two **[18]**. VRD can be used as the main information for image captioning and is an independently essential information.

Therefore, to solve this problem, we propose a Context Pair Network (CPN) that describes only the context of an object based on VRD. The proposed model consists of a pair object module (POM) that extracts subjects, objects, and relationships and a pair embedding module (PEM) that creates a subject–predicate–object structure. The proposed POM detects objects corresponding to subjects and objects based on three RCNNs and matches the detected objects with subject–object (S-O) pairs. The PEM also generates subject–predicate–object sentences using S-O pairs based on LSTM. Thus, the proposed model is capable of situational awareness that provides only interactions between objects, and not conventional captioning, which describes all information indiscriminately. Experiments were conducted using the Visual Genome (VG) dataset to verify the proposed model **[16]**.

2. Related studies

2.1 Encoders in image captioning

Image captioning is divided into two steps: an encoder that extracts a specific vector from an image and a decoder that generates sentences. In this structure, the feature vector of the image containing the object is required, not the class information of the object. Therefore, because image captions generally have an object recognition model as an encoder, only the previous feature vector is used without using the result of the classifier of the model. Thus, the expressive power of the sentence is improved according to the performance of the encoder. Therefore, to improve the accuracy of image captioning, a model that can accurately classify various objects must be selected.

As a representative object recognition challenge, the ImageNet Large Scale Visual Recognition Competition (IRSVRC) introduced excellent deep learning models. In particular, the models of VGGNet **[9]**, ResNet **[10]**, Inception v3, and Inception v4 **[11][12]** are primarily applied as encoders for image captions. The NIC model was constructed based on basic VGGNet. ResNet solves the gradient loss problem caused by the deepening of CNN layers through residual blocks and has been widely applied to image caption models. Object detection models such as RCNN and YOLO, which have relatively complex structures, have been

applied to a limited extent compared with existing object recognition models owing to their object classification types and computational efficiency. However, an image caption model that uses an object detection model capable of detecting small and various objects has been studied. Object detection models, such as YOLO **[13][14]** and faster-RCNN **[15]**, are typically used as encoders. In particular, faster-RCNN has been widely used as an encoder because it can detect more precisely than YOLO and is relatively easy to modify. However, faster-RCNN has a disadvantage because the detection speed is relatively slow compared with that of YOLO.

2.2 Dense image captioning

Existing image captions provide concentrated information, because only one sentence is obtained from a single image. However, this approach results in loss of local information in surveillance systems that capture a wide area. Hence, dense image captions acquire several captions from a single image. Therefore, it is necessary to provide a plurality of object information to the decoder during the encoder step. However, the trained model has limited capabilities, because the conventional approach feeds one feature vector to the decoder for the entire image. Methods for generating and inputting feature vectors by separating each object unit from an image have been studied to solve this problem.

Johnson *et al.* **[1]** applied an object detection model that outputs the location information of an object to input the feature vectors of various objects present in the image. This FCLN extracts a feature vector corresponding to each object region by estimating the relative position of the image feature vector based on the position of the bounding box presented in faster-RCNN. This model performs individual captions based on the extracted feature vectors, thereby enabling captioning for each object.

Yang *et al.* **[7]** and Li *et al.* **[8]** separately analyzed the context features from the entire image to obtain the background and context information lost owing to feature extraction. This method has the advantage of being able to describe the interaction between an object and the surrounding space; however, it still has limitations in outputting only local information. This limitation occurs because only extremely local information is provided to the decoder. Finally, only one object was acquired when feature vectors were added. To solve this problem, information on complex object clusters must be provided.

2.3 Visual relationship detection

VRD is a technique for detecting the position of a pair of objects that have a relationship with each other among the objects existing in an image and the relationship between them. VRD defines the relationship between two objects and one predicate that connects them, expressed in the form of <subject predicate, object>, which is called a triplet. The relationship between objects is Action, Spatial, Preposition, and Comparative Verbs are classified into a total of five, and each object is classified into a class through an object detection model.

Lu *et al.* **[17]** proposed an efficient relationship detection model by learning visual information such as external features of objects and the classification and relationship information of each object composed of natural language independently of each other. Zhan *et al.* **[18]** proposed a new network that can improve performance by simultaneously using unlabeled data for learning to improve the VRD performance in scenarios where the amount of data is insufficient. Recently, graph-based models have been proposed to extract object-level information by representing objects as nodes and predicates as edges by applying a graph structure to VRD **[19]**. In particular, Mi *et al.* **[19]** proposed an attention-based graph model that predicts relationships using the dependencies between each triplet rather than the correlation between objects constituting the triplet. Thus, the model enabled the creation of a triplet-level graph rather than an object-level graph.

However, because the VRD model outputs only the relationship between objects in words, the amount of information that can be acquired is limited. In addition, the model cannot output results for unlabeled objects. In addition, while VRD outputs a relationship for all cases in which detected objects are paired, there is a limit in that the detection accuracy is relatively low because not all objects have a relationship. Therefore, VRD outputs information about an object as a sentence, as in image captioning. It must be combined with technology that can explain unlabeled data to compensate for its disadvantages.

3. Proposed dense image captioning

3.1 Overview of the proposed method

A completely unmanned surveillance system must be informed of all objects within the images acquired by CCTV. Dense image captioning creates a description for each detected object. However, this requires more information such as the color or position of the object. As a result, even detailed information about the

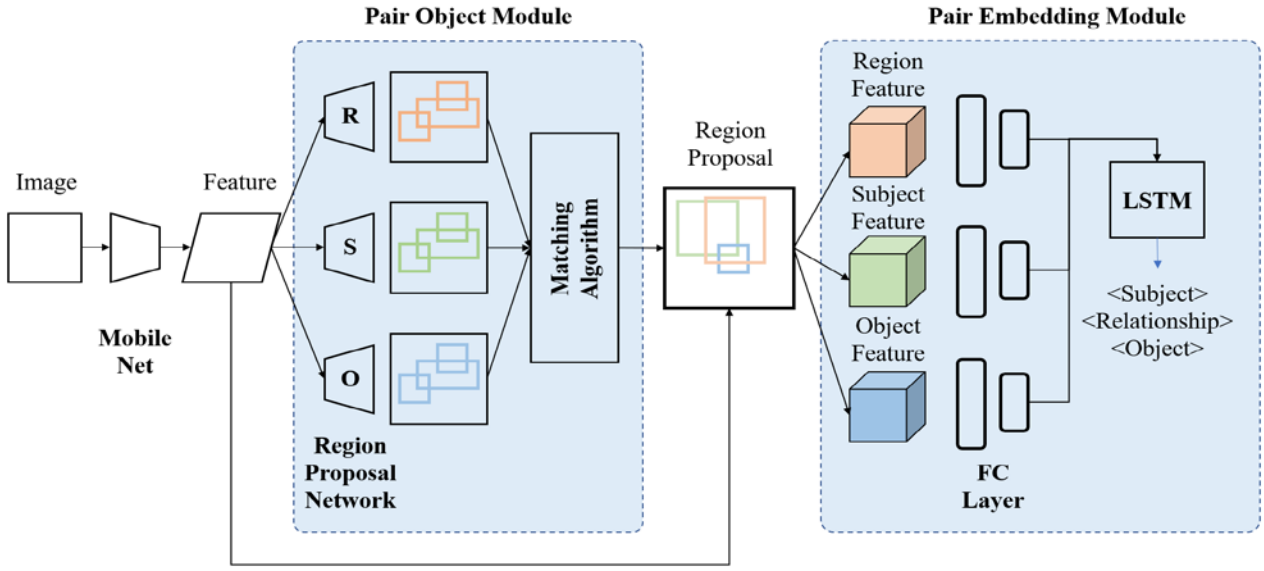

Figure 1: Architecture of the proposed context pair network

target object can be expressed, but only the surrounding information is expressed, and not between objects. In this study, to solve this problem, we propose a CPN combined with VRD that outputs vital elements rather than generating descriptive sentences.

People generally require a subject, predicate, and object to describe a scenario. VRD includes relationships corresponding to this subject, object, and predicate. VRD can describe the scenario in its simplest form. The proposed CPN includes a POM that extracts subject, relation, and object information by combining VRD and a PEM that creates a subject–predicate–object structure. **Figure 1** shows the structure of the proposed model.

The proposed model uses MobileNet **[24]** to acquire feature vectors in the images. Algorithms with multiple stages, including areas of VRD, use well-learned object recognition models to easily acquire feature vectors in the images. This recognition model is known as the backbone. As previously mentioned, the performance of the model is highly dependent on the performance of the backbone. However, modern object recognition models based on intense networks exhibit strong performance. However, classes are not required because VRD uses a separate language model. Therefore, the proposed model requires a lightweight backbone even if it is inaccurate. Therefore, MobileNet was selected in this study.

The two blue boxes in **Figure 1** are POM and PEM, respectively. The proposed POM detects the location of a sentence, region, subject, and object based on three faster-RCNNs. Moreover, the POM matches the proper subject and object. The region proposals derived through this process are pooled as feature vectors through the region proposal network (RPN). The PEM then projects this feature vector using a fully connected (FC) layer and generates a subject–predicate–object sentence composed through LSTM. Thus, the relationship between the subject and object can be described.

3.2 Proposed the pair object module

The existing VRD detects the area of the relationship by finding the union of the two, including the subject and object. These unions are difficult to learn because they do not exist in real datasets. To solve this problem, most studies have used a method of matching GT first by using the characteristics between the S-O pair and the region, which is the region of a sentence. Although this approach is intuitive and logically correct, it has limitations owing to the need for datasets to learn it. In addition, because object detection models such as the RCNN series are used separately, detecting all objects and individually matching them is required.

The proposed POM uses two tasks to recognize subjects, objects, and domains while reducing these uncertainties. First, the module minimizes the size of the detection area. It simultaneously increases clarity by detecting the region that is the object of description rather than the union. Second, the accuracy of the S-O pair is increased by matching the detected subject and object based on the region. These two processes are the RPN and matching algorithm in **Figure 1**.

The proposed POM first inputs feature vectors processed in

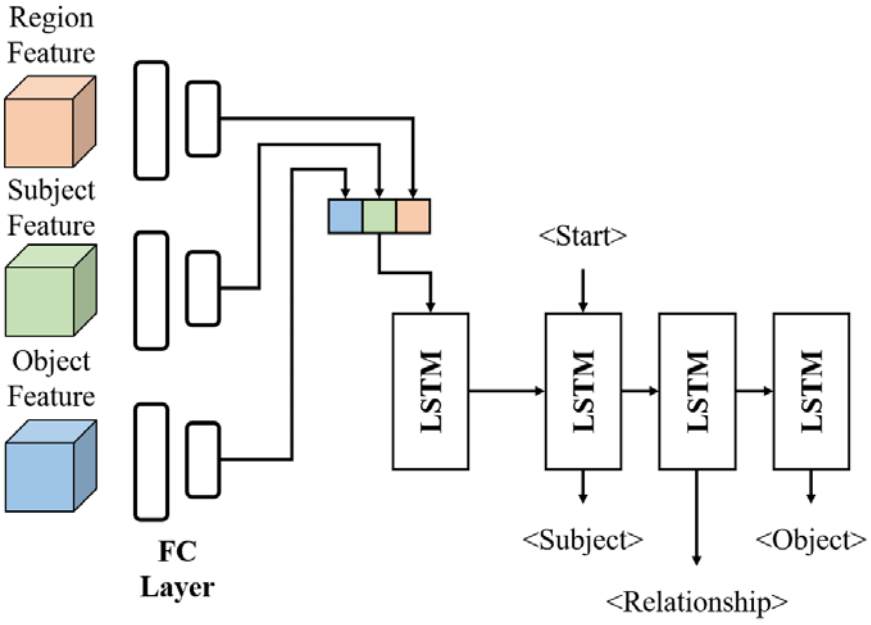

Figure 2: Structuue of the pair embedding module

the backbone into three pairs of RPNs. It comprises an RPN that detects regions and an RPN that detects subjects and objects. These three networks have the same structure but independent weights. Unlike the RPNs of S and O, which detect objects, R detects areas; therefore, its accuracy is relatively low and its size is large. These three networks learn through multi-task learning.

Because the detected S-O and R are independent of each other, matching is essential. The proposed POM selects as many candidates as the number of cases in which all the detected S's and O's could be matched. Furthermore, it creates a union between the candidates. The POM matching algorithm analyzes unions and regions to determine S-O pairs. This matching process identifies the degree of overlap through the intersection over union (IOU) and filters through a threshold. Unlike conventional methods, this method maintains the form of a dataset that generates actual sentences and simultaneously inputs S-O pairs.

3.3 Pair embedding module

Basic VRD transforms an RCNN's RPN to classify subjects, objects, and relationships, such as object recognition. However, object recognition classes can represent multiple objects as single representative objects. However, the number of words used for text is large because of derivatives such as synonyms. Therefore, it is difficult to classify them into classes. In the proposed PEM, a language model is separately trained using LSTM to solve this problem. Through this, more diverse words can be learned compared with existing methods, and even exceptional cases, such as derivatives, can be covered.

Figure 2 shows the structure of the proposed PEM. The POM receives the features pooled from POM as input, as shown in Figure 1. Because networks that receive feature vectors from other models have different dimensions at the front and back, it is necessary to match them. In particular, it is essential in the case for a non-dimensional layer, such as LSTM or FC. Because the POM has three feature vectors, they must be matched to the same size. Therefore, a projection network composed of FC layers is typically inserted. This network does not simply match the size but also normalizes the results to transform them such that they can be processed in a later network. In the proposed POM, the size is changed to 512 through two FC layers and concatenated with each other.

The decoder that creates the sentence uses LSTM. Techniques such as image captioning, which describes visual features, use an initial injection method to preemptively process feature vectors before generating sentences through LSTM. The initial injection m ethod preprocesses and conceals the feature vector through one LSTM cell before inputting '<Start>,' a token that starts a sentence. Recently, the parallel injection method, which inputs the input at the output moment of each word, has been widely used. The parallel injection method is more robust against memory vanishing, which occurs in a longer sequence than in the initial injection method. However, because the PEM outputs short three-word sentences, the memory vanishing problem does not

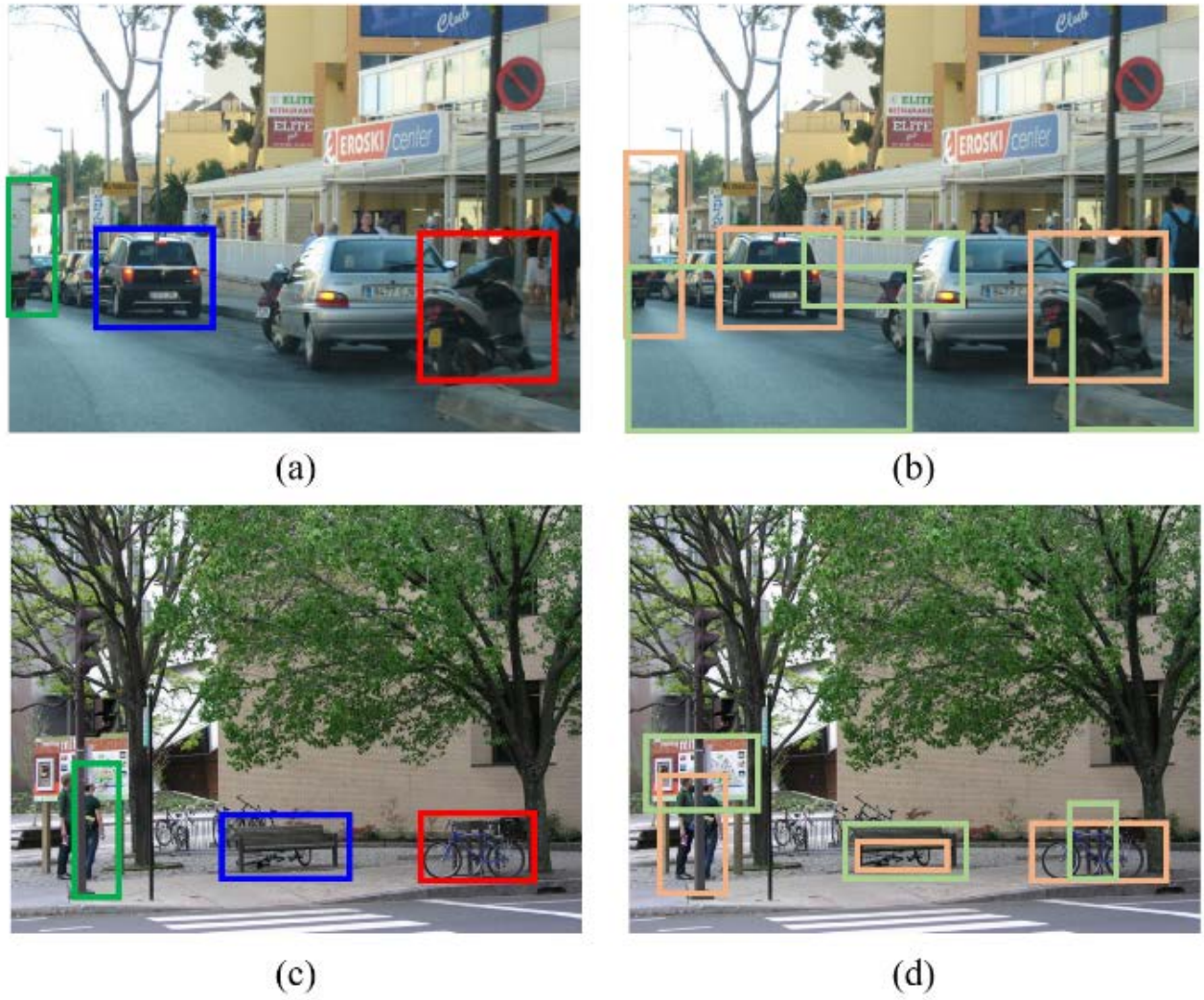

Figure 3: The result of proposed model

occur. Therefore, this module uses a relatively lightweight initial injection method.

4. Experiment and result

4.1 Datasets and preprocessing

This study verified the proposed model using the VG dataset **[16]**. The VG dataset was built to extract more complex information from images beyond simple object recognition, such as dense image capture, visual question answering (VQA), and region graphs. Three versions of the VG dataset are currently open: VG 1.0, VG 1.2, and VG1.4. In this study, VG 1.2 was used for training and verification. The VG 1.2 dataset contains 108,249 images, approximately 4,100,000 sentences, and 2,300,000 relational data. Because not all image regions have a relationship, only regions with a relationship attribute are extracted first among all regions. Each region includes one sentence, the subject and object that compose the relationship, and their relationship information. A total of 85,200 images remaining after this process were used for learning and verification. The training and validation test data classification was assigned to 75,456 training, 4,871 validation, and 4,873 test data, as presented in **[1]**.

All processes of the proposed model were conducted in Python, and the network was implemented through the PyTorch framework.

4.2 Experiment and result

The proposed model detects the subject, object, and region through the POM to recognize the context of the image and creates a subject–predicate–object structure through the PEM. This model aims to create appropriate positions and pairs of subjects and objects and generate correct sentences. Unlike general VRD, the proposed model cannot use BLEU, which compares the

matching of consecutive words. Therefore, in this study, we demonstrated the performance by qualitatively comparing the mAP, which is the matching accuracy of the bounding box and sentences.

Figure 3 shows the S-O pairs detected using the GT and POM. (a) and (c) in **Figure 3** represent the GT, and (b) and (d) show the results. Because the GT is a box in a region, only one box was created. Therefore, two objects were detected in each area, and could be considered a subject and an object, respectively. In **Figure 3**, we observe that two S-O pairs were matched in all areas. In particular, in **Figure 3(b)**, the green box was captured as the background, not the object. This phenomenon occurred because the dataset not only treats objects as objects in sentences but also uses features such as roads and buildings as objects. Because of these characteristics, the model can provide more detailed descriptions. Moreover, in **Figure 3(d)**, only the surrounding objects were designated neatly. We can qualitatively confirm that the S-O bond was appropriately matched in the actual image, as shown in **Figure 3**.

Table 1 compares the evaluation indices of each model trained with the VG dataset to compare the performance of the proposed model with recent dense image captioning models. Dense image capturing models are designed for different purposes, and the preprocessing process still requires to be unified. Therefore, simple numerical comparisons differ depending on the number of classes or size of the matched data. The mAP scores in **Table 1** show that the proposed model obtained the highest score (11.3 %).

Table 1: Results of mAP score by models on the dataset

Model	mAP (%)
FCLN [1]	5.39
CAG-NET [21]	10.51
T-LSTM [7]	9.96
COCG [8]	8.90
Proposed model	**11.3**

5. Conclusion

In this paper, we propose a CPN that describes only the interactions between objects without indiscriminately describing all information. The proposed model consists of a POM that detects subjects, objects, and regions and a PEM that generates them as sentences. Using this model, captioning that describes only the scenario of an object is possible by improving the existing dense image captioning, which provides a large amount of unnecessary information. The performance of the proposed model was compared and evaluated using the mAP. The mAP exhibited a high performance of 11.3%.

Through this study, we expect that relationships between objects can be explained more accurately with expressive sentences in the field of dense captions in the future.

Acknowledgments

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). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C1014024).

This research was supported by Korea Basic Science Institute (National research Facilities and Equipment Center) grant funded by Ministry of Education(grant No. 2022R1A6C101B738).

Author Contributions

Conceptualization, J. H. Seong; Methodology, J. H. Seong; Software, Y. J. Shin; Formal Analysis, Y. J. Shin and S. B. Jeong; Investigation, Y. J. Shin; Resources, J. H. Seong; Data Curation J. H. Seong; Writing-Original Draft Preparation, Y. J. Shin and S. B. Jeong; Writing-Review & Editing, J. H. Seong; Visualization, Y. J. Shin; Supervision, D. H. Seo; Project Administration, D. H. Seo; Funding Acquisition, D. H. Seo.

References

[1] J. Johnson, A. Karpathy, and L. Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4565-4574, 2016

[2] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” European conference on computer vision, Springer, Berlin, Heidelberg, pp. 15-29. 2010.

[3] Y. Yang, C. Teo III, H, Daumé, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 444-454, 2011.

[4] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-

grams," Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220-228, 2011.

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

[6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651-4659, 2016.

[7] 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. 1978-1987, 2017.

[8] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 8650-8657, 2019.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556., 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826, 2016.

[12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," Thirty-first AAAI Conference on Artificial Intelligence, vol. 31, no. 1, 2017

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi "You only look once: Unified, real-time object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[14] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[15] 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.

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol. 123, no. 1, pp. 32-73, 2017.

[17] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," European Conference on Computer Vision, pp. 852-869, 2016.

[18] Y. Zhan, J. Yu, T. Yu, and D. Tao, "On exploring undetermined relationships for visual relationship detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5123-5132, 2019.

[19] L. Mi and Z. Chen, "Hierarchical graph attention network for visual relationship detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13883-13892, 2020.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861. 2017.

[21] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6234-6243, 2019.