Sensors and Materials, Vol. 35, No. 7 (2023) 2195–2204
MYU Tokyo

S & M 3316

Image Caption Generation Using Scoring Based on Object Detection and Word2Vec

Tadanobu Misawa,[1*] Nozomi Morizumi,[2] and Kazuya Yamashita[3]

[1]Faculty of Engineering, University of Toyama,
3190, Gofuku, Toyama-shi, Toyama 930-8555, Japan
[2]Graduate School of Science and Engineering for Education, University of Toyama,
3190, Gofuku, Toyama-shi, Toyama 930-8555, Japan
[3]Faculty of Education and Research Promotion, University of Toyama,
3190, Gofuku, Toyama-shi, Toyama 930-8555, Japan

(Received March30, 2023; accepted June 6, 2023)

Keywords: image caption generation, deep learning, object detection, Word2Vec, scoring

Generating descriptive text from images, known as caption generation, is a noteworthy research field with potential applications, including aiding the visually impaired. Recently, numerous methods based on deep learning have been proposed. Previous methods learn the relationship between image features and captions on a large dataset of image–caption pairs. However, it is difficult to correctly learn all objects, object attributes, and relationships between objects. Therefore, occasionally incorrect captions are generated. For instance, captions about objects not included in the image are generated. In this study, we propose a scoring method using object detection and Word2Vec to output the correct caption for an object in the image. First, multiple captions are generated. Subsequently, object detection is performed, and the score is calculated using the resulting labels from object detection and the nouns extracted from each caption. Finally, the output is the caption with the highest score. Experimental evaluation of the proposed method on the Microsoft Common Objects in Context (MSCOCO) dataset demonstrates that the proposed method is effective in improving the accuracy of caption generation.

1. Introduction

Recently, there has been a surge in research into artificial intelligence, particularly deep learning, with numerous practical applications. Starting from image recognition, it has developed into image generation using generative adversarial networks. Additionally, research is also being conducted on automatic generation of explanatory text for images, such as image caption generation.[1–5] Thus, image caption generation can generate text from visual information (e.g., images) obtained from a camera or other sensors and convert it into information that can be confirmed by listening using the read-aloud function. Expectations are that the quality of life of visually impaired people will be improved.

*Corresponding author: e-mail: misawa@eng.u-toyama.ac.jp
https://doi.org/10.18494/SAM4410

ISSN 0914-4935
https://myukk.org/

Vinyals *et al.* proposed a technique with image feature extraction and caption generation modules.[1] First, image features were extracted using Convolutional Neural Networks (CNNs). Subsequently, a Long Short-Term Memory (LSTM) network generates captions on the basis of image features. Numerous other studies have made use of such methods using CNN and LSTM, a common approach of which is described in a study by Xu *et al.*[2] In their study, attention was added to the method; more specifically, image features were extracted from the input image using a CNN, and then a caption was generated using attention and LSTM. Thus far, numerous attention-based methods have been proposed.

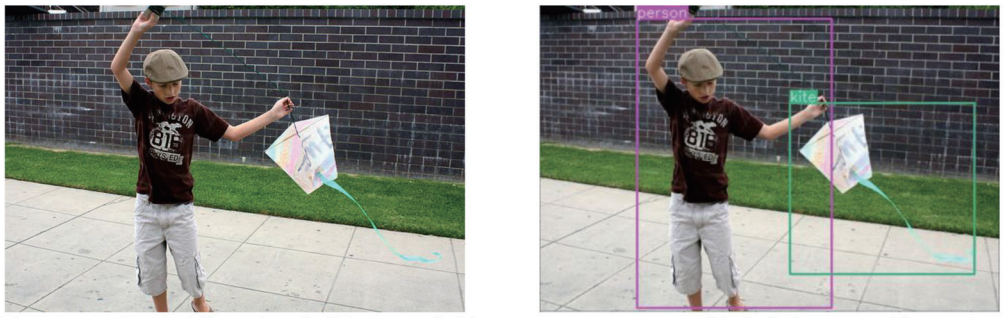
Such a caption generation model is trained on a large dataset of image–caption pairs. Alternatively, it learns the relationship between image features and teacher labels. However, correct learning for all images is difficult as the recognition of objects in the image and their attributes, behaviors, and relationships between objects must be learned. Therefore, incorrect captions may be generated. To improve the accuracy of caption generation, extracting a large amount of information from images is important. Therefore, research has been conducted on object detection methods for extracting object information from images. Object detection can extract the position (as a rectangle) and label (person, dog, etc.) of an object. In the study by Li *et al.*,[3] the object regions extracted by object detection were input to a CNN to obtain object image features. The features of the entire image and each object were adjusted by attention and used as image features. Iwamura *et al.* reported a technique that integrated object detection and motion estimation to extract motion information from images.[4] In addition, methods using object regions and labels have been proposed. Baig *et al.* reported a technique that generated a caption and then replaced the words in the caption with the extracted label using object detection.[5]

In this study, we propose a method for scoring captions by focusing on the labels extracted by object detection. Specifically, multiple captions are generated, and these captions are scored using Word2Vec with the object nouns in each caption and labels extracted via object detection. Finally, the output is the caption with the best score, which should enable the output of correct captions about objects in the image.

2. Data, Materials, and Methods

It is possible to generate captions for images not existing in the training data using deep learning techniques such as CNN and LSTM. However, it is difficult to accurately identify objects and describe relationships between them in many images. Therefore, image caption generation may fail by outputting descriptions of objects that do not appear in the image. However, object detection may correctly extract object labels. Figure 1 illustrates an example of caption generation and object detection. Caption generation outputs the incorrect noun, “baseball bat,” but object detection correctly extracts “kite.”

In our study, we first generate multiple captions. Subsequently, we propose a method to output captions that correctly recognize objects in the image by scoring them based on semantic proximity, which is the cosine similarity using Word2Vec.



(a) a young boy holding a baseball bat on a field (b) Label: person, kite

Fig. 1. (Color online) An example of (a) caption generation and (b) object detection.

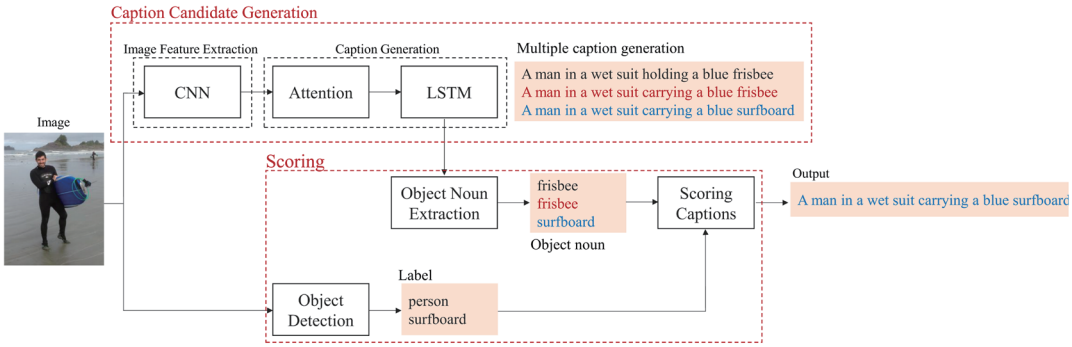


Fig. 2. (Color online) Overview of proposed method.

2.1 Overview

The overall scheme of the proposed method is shown in Fig. 2. The proposed method comprises two stages: caption candidate generation and scoring. First, in the caption candidate generation module, multiple captions are generated from an image using a deep learning model and the beam search algorithm. Subsequently, each generated caption is scored in the scoring module, and finally the caption with the highest score is outputted. The score of each caption is calculated on the basis of the semantic similarity between the object nouns in the caption and the labels obtained by object detection. Therefore, the score of the generated caption that contains an incorrect object noun, as in the example shown in Fig. 1, is lower. Finally, the output of the caption with the higher score should improve the accuracy of the caption generation.

2.2 Caption candidate generation

In our study, multiple captions were generated on the basis of the method proposed by Xu *et al.*[2] That is, image features were extracted from the image using a pretrained CNN, and then weighted image features were generated by the attention mechanism. The LSTM then generated

captions on the basis of these features. We used the pretrained ResNet152[6] to extract image features. ResNet is a model that offers high accuracy in image classification tasks and is also used in image caption generation. In the LSTM used for multiple caption generation, the hidden layer and the distributed representation of words had 512 dimensions. For training, a cross-entropy error was used as the loss function, and Adam was used as the optimization algorithm. The batch size was set to 128, and training was terminated when the minimum cross-entropy error could not be updated for five consecutive epochs. When generating multiple captions, the beam search algorithm, which is also used in machine translation, was used to generate multiple captions based on the beam width. In Fig. 2, when the beam width is equal to three, the caption candidates generated are "A man in a wet suit holding a blue frisbee", "A man in a wet suit carrying a blue frisbee", and "A man in a wet suit carrying a blue surfboard".

2.3 Object detection

Object detection is a method used to extract objects from images, and the output is typically the region of the object in the image and its label. In numerous previous methods of image caption generation, object regions detected from an image were often input to the CNN and used as object image features. Therefore, we focus on labels rather than object regions. Figure 2 shows that "person" and "surfboard" have been detected via object detection, and these labels are used. In this study, we used the pretrained Faster R-CNN[7] for object detection, accepting objects with a confidence score of 0.5 or higher.

2.4 Semantic similarity using Word2Vec

We calculated the semantic similarity between words using cosine similarity, where each word was converted into a vector using Word2Vec.[8] Word2Vec is a method used to obtain a distributed representation of words using a neural network. In this study, GoogleNews-vectors-negative300 was used as a model for Word2Vec.

2.5 Object noun extraction

Nouns were extracted from the generated caption candidates. Here, non-object nouns such as "road", "street", and "park" are included. In this study, the object detected is used as the basis for scoring the caption, such that nouns that do not imply objects must be excluded from the score calculation. Therefore, we extracted object nouns with a semantic similarity of λ or more to one of all possible labels for object detection. In Fig. 2, "man", one of the nouns extracted from the caption candidates, is not extracted as an object noun because its semantic similarity to all possible labels for object detection is below the threshold λ, whereas "frisbee" and "surfboard" were extracted as object nouns because their semantic similarity to all possible labels for object detection exceeded the threshold λ.

2.6 Scoring caption

The semantic similarity between each object noun extracted from the candidate captions and each label obtained by object detection is calculated. Subsequently, the maximum value is considered the score of each object noun, and the score of each candidate caption is the average score of all object nouns in that caption. There are several captions for which no score can be calculated (no object nouns) because the nouns used to calculate the score have been restricted to exclude nouns that are not objects. In such cases, the caption score is set to zero.

An example of the scoring caption is shown in Fig. 3. In Fig. 3, the labels "oven" and "pizza" are obtained by object detection. Moreover, "stove" and "pizza" are extracted as object nouns from the caption candidate "a close up of a pizza pan on a stove". The cosine similarity between the object noun "stove" and the labels "oven" and "pizza" is calculated. As a result, the cosine similarities are 0.6 and 0.2, respectively, and the same is calculated for the object noun "pizza". Subsequently, the maximum value of the cosine similarity in each object noun is taken as the object noun score (0.6 for "stove" and 1.0 for "pizza"), and the average of all object noun scores (0.8 in Fig. 3) is the caption candidate score.

2.7 Final caption

From the results of the caption scoring, the output is the caption with the highest score. For instance, in Fig. 2, "A man in a wet suit holding a blue frisbee" is output when no scoring is performed. However, when a caption is scored on the basis of the semantic similarity of the object nouns "frisbee", "frisbee", and "surfboard" of each caption candidate and the labels "person" and "surfboard" obtained through object detection, the highest scored caption is the caption with "surfboard". Therefore, "A man in a wet suit carrying a blue surfboard" is the final output caption. Thus, by scoring captions, accurately worded captions are generated.

3. Results

Experiments were conducted to validate the effectiveness of the proposed method. A large dataset, Microsoft Common Objects in Context (MSCOCO),[9] was used. One hundred thirteen thousand two hundred eighty-seven images were used for training, 5000 for validation, and 5000

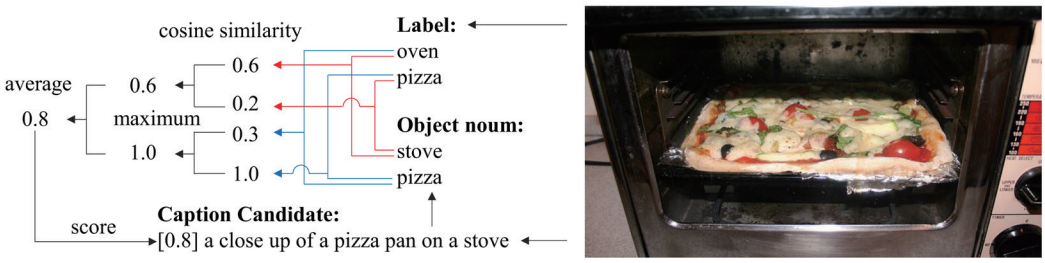


Fig. 3. (Color online) An example of the scoring caption.

for testing. The beam width for generating multiple captions as candidate captions was set to three, and the threshold λ for extracting object nouns for scoring captions was set to 0.6, because the best results were obtained by experimenting with a threshold λ from 0.40 to 0.80 in increments of 0.05. These results suggest that a small threshold λ extracted object nouns that had a weaker relationship with object detection labels, whereas a large threshold λ extracted only object nouns that had a stronger relationship with object detection labels than necessary, resulting in unsuitable score calculation.

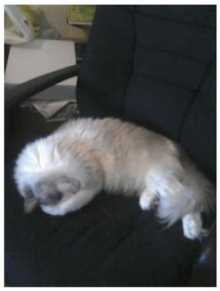
BLEU,[10] METEOR,[11] ROUGE-L,[12] and CIDEr[13] were used as evaluation indices. Higher ratings indicate greater accuracy of captions. BLEU is an evaluation index based on the percentage of n-gram agreement between generated and supervised captions. METEOR is an evaluation index based on the percentage of word agreement between generated and supervised captions. ROUGE-L is an evaluation index based on the longest common subsequence. CIDEr is a proposed metric for evaluating caption generation that considers the TF-IDF weight of the number of n-gram occurrences.

The experimental results of the baseline [previous method (2)] and proposed methods are shown in Table 1. From Table 1, a slight improvement in evaluation is observed compared with the baseline method, except for METEOR. The proposed method generated different captions from the baseline method for 1090 images, and the accuracy of the captions of these images may be improved because of the inclusion of more accurate words than those obtained by the baseline method.

The specific caption scoring results are shown in Fig. 4. Figure 4 shows the target image, the labels extracted through object detection, and the generated caption candidates and their scores;

Table 1
Results of baseline and proposed methods with MSCOCO dataset.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Baseline	70.3	53.2	39.6	29.6	24.6	52.0	93.6
Proposed	70.7	53.6	39.9	29.8	24.6	52.1	94.3



Label:
cat, couch
Generated caption:
[0.6628656] a white cat laying on top of a suitcase
[0.7119198] a white cat laying on top of a blue chair
[0.6418425] a white cat laying on top of a blue bag

(a)

Label:
ovens, person, dog, bottle, oven
Generated caption:
[0.0] a man holding a small child in a cage
[0.0] a man holding a small child in a kitchen
[0.76094574] a man holding a cat in a kitchen

(b)

Fig. 4. (Color online) Examples of generated captions and scores.

the underlined words in the captions are the object nouns used for the score calculation. Figure 4(a) shows an image of a cat lying on a chair; the "cat" and "couch" are extracted through object detection. The scores of each caption differ according to "suitcase," "chair," and "bag," and the caption containing a "chair," which is closest to the label "couch," has the highest score. Therefore, without scoring, the caption including the suitcase is outputted. However, the proposed method can output the correct caption including chair even if the chair is not extracted by object detection. In the caption generation in Fig. 4(b), the score of two captions is 0.0 because their scores are not calculated. In object detection, all objects related to people are recognized as labeled "person." However, the nouns "man" and "child" extracted from the captions are not used in the score calculation because their semantic similarity to all detected object labels, including "person," is below the threshold $\lambda = 0.6$. Furthermore, object detection extracted a dog rather than a cat. However, the score of the third caption was calculated, and this caption was output because the cat and the dog are semantically close.

Figure 5 compares the output captions in the baseline and proposed methods. Figures 5(a) and 5(b) are examples of improved captioning compared with that of the baseline method, and Figs. 5(c) and 5(d) are examples of poor captioning compared with that of the baseline method. Figures 5(a) and 5(b) demonstrate that the baseline method outputs captions that include "frisbee" and "tennis racket," which are not in the target image. In contrast, the proposed method outputs correct captions that include "surfboard" and "toothbrush." However, as shown in Fig. 5(c), when the correct caption includes a noun such as baseball, which is not an object, the proposed method outputs an incorrect caption. Furthermore, as shown in Fig. 5(d), if the object detection fails to recognize a guitar correctly and objects that are semantically close to a guitar cannot be recognized, an incorrect caption will be output.The proposed approach is to change the output caption if there is a more accurate caption for an object noun in the image among the multiple captions generated by the baseline method. Alternatively, the proposed method improved accuracy by postprocessing after generating multiple captions via the previous method. Therefore, the proposed method can be applied to various existing methods to improve their accuracy.

4. Discussion

Since the proposed method scores caption candidates, generating more caption candidates (i.e., increasing the beam width) can be expected to improve the accuracy. Therefore, Table 2 shows the results when the beam width is increased. In Table 2, it can be seen that as the beam width increased, the accuracy decreased. It has been reported that when the beam width increases, the search space becomes more extensive, but the accuracy does not improve.[14] The same tendency was observed for the proposed method.

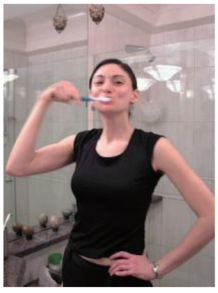
Therefore, we further investigated if increasing the number of caption candidates using multiple small beam widths improved accuracy. For example, if caption candidates were generated using a combination of beam widths from 1 to 3, six caption candidates can be generated. The results for beam width combinations of 1 to 3, 1 to 4, and 1 to 5 are shown in Table 3. Table 3 shows the best results for beam width combinations of 1 to 3, indicating that the

Label:
surfboards, person, surfboard, people

Baseline:
a man in a wet suit holding a blue frisbee

Proposed method:
a man in a wet suit carrying a blue surfboard

(a)



Label:
sink, vases, toothbrushes, base person, toothbrush, bottle

Baseline:
a woman holding a tennis racket in her hand

Proposed method:
a woman in a white shirt holding a toothbrush

(b)

Label:
person, chair, people

Baseline:
a group of young men playing a game of baseball

Proposed method:
a group of young men playing a game of frisbee

(c)

Label:
handbag, book, beds, bottles, bed, books, person, bottle

Baseline:
a young man sitting on a couch with a guitar

Proposed method:
a young man sitting on a couch with a remote

(d)

Fig. 5. (Color online) Examples of captions generated by baseline and proposed methods.

Table 2
Results of proposed method with different beam widths.

Beam width	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
3	70.7	53.6	39.9	29.8	24.6	52.1	94.3
5	68.6	51.3	38.0	28.2	23.8	50.6	91.2
10	66.0	48.5	35.3	25.8	22.6	48.5	84.3
20	92.8	45.2	32.4	23.5	20.8	45.8	76.6

Table 3
Results of proposed method with multiple beam widths.

Beam width	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
1,2,3	71.1	54.1	40.3	30.0	24.7	52.4	95.8
1,2,3,4	71.1	53.9	40.2	30.0	24.6	52.2	95.5
1,2,3,4,5	71.0	53.8	40.1	29.9	24.6	52.1	95.2

accuracy was further improved. In particular, the improvement in the accuracy of the evaluation index CIDEr is significant. The CIDEr score is a TF-IDF-based score, which is considered to be strongly affected by the importance of the words. Therefore, the accuracy of CIDEr is improved by outputting a caption that contains accurate words. The numbers of caption candidates generated for beam width combinations from 1 to 4 and beam width 10 are the same, but the accuracies are significantly different. Therefore, when increasing the number of caption candidates, combining several smaller beam widths is found to be more effective than increasing the beam width.

5. Conclusions

In this study, we proposed an improved method for including objects that do not appear in the image when generating image captions. Specifically, multiple captions were generated by the beam search algorithm on the basis of existing methods, and nouns were extracted from each caption. Because our study focuses on objects in the image, a threshold λ was introduced to extract object nouns from the nouns in each caption. Subsequently, object detection was used to detect labels of objects in the image. Moreover, the score of each caption is calculated using the object nouns extracted from each caption, and the object detected labels by cosine similarity using Word2Vec. Finally, the caption with the highest score was outputted.

Experimental results demonstrated that the proposed method was effective in improving the accuracy of image caption generation. In addition, because the proposed method improved the accuracy by postprocessing, it can be applied to various existing methods. Improved accuracy in image caption generation is anticipated to advance information services based on sensors such as cameras and improve our lives. In future work, it is necessary to improve the accuracy of object detection and handling of nouns other than object nouns to improve the accuracy of image caption generation.

References

1 O. Vinyals, A. Toshev, S. Bengio, and D. Erhan: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2015) 3156.
2 K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio: Proc. 32nd Int. Conf. Machine Learning (PMLR 37, 2015) 2048.
3 L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian: Proc. AAAI Conf. Artificial Intelligence **31** (2017). https://doi.org/10.1609/aaai.v31i1.11236
4 K. Iwamura, J. Y. L. Kasahara, A. Moro, A. Yamashita, and H. Asama: Sensors **21** (2021) 1270. https://doi.org/10.3390/s21041270
5 M. M. A. Baig, M. I. Shah, M. A. Wajahat, N. Zafar, and O. Arif: Proc. 2018 Digital Image Computing: Techniques and Applications (DICTA, 2018). https://doi.org/10.1109/DICTA.2018.8615810
6 K. He, X. Zhang, S. Ren, and J. Sun: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016) 770.
7 S. Ren, K. He, R. Girshick, and J. Sun: Adv. Neural Inf. Process. Syst. **28** (2015).
8 T. Mikolov, K. Chen, G. Corrado, and J. Dean: arXiv preprint at arXiv:1301.3781 (2013). https://doi.org/10.48550/arXiv.1301.3781
9 T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick: Proc. European Conf. Computer Vision (2014) 740. https://doi.org/10.1007/978-3-319-10602-1_48
10 K. Papineni, S. Roukos, T. Ward, and W. J. Zhu: Proc. 40th Annu. Meeting of the Association for Computational Linguistics (ACL, 2002) 311. https://doi.org/10.3115/1073083.1073135

11 M. Denkowski and A. Lavie: Proc. 9th Workshop on Statistical Machine Translation (2014) 376. https://doi.org/10.3115/v1/W14-3348
12 C. Y. Lin, G. Cao, J. Gao, and J. Y. Nie: Proc. Human Language Technology Conf. the North American Chapter of the ACL (2006) 463.
13 R. Vedantam, C. L. Zitnick, and D. Parikh: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2015) 4566.
14 O. Vinyals, A. Toshev, S. Bengio, and D. Erhan: IEEE Trans. Pattern Anal. Mach. Intell. **39** (2016) 652. https://doi.org/10.1109/TPAMI.2016.2587640

About the Authors

Tadanobu Misawa received his B.E., M.E., and Ph.D. degrees from Kanazawa University in 1999, 2001, and 2004, respectively. He was a doctoral fellow at the Kanazawa Institute of Technology in 2004. From 2005 to 2008, he was an assistant professor at the School of Management, Tokyo University of Science. He is currently an associate professor at the Faculty of Engineering, University of Toyama. His research interests are artificial intelligence and brain–computer interface. (misawa@eng.u-toyama.ac.jp)

Nozomi Morizumi received his B.E. degree in Intellectual Information Engineering from the University of Toyama, Toyama, Japan, in 2021. Currently, he is a graduate student at the Graduate School of Science and Engineering for Education, University of Toyama. His research interest is image caption generation.

Kazuya Yamashita received B.E., M.E., and Ph.D. degrees in engineering from the University of Toyama in 2004, 2006, and 2018, respectively. He was a technical staff at the Faculty of Engineering, University of Toyama in 2006. He was an assistant at the Faculty of Engineering, University of Toyama, from 2007 to 2017. He has been a lecturer at the Information Technology Center, University of Toyama, since 2018. His research interests are formal language theory, automata theory, and algorithm theory. (kazuya@itc.u-toyama.ac.jp)