ICIC Express Letters
Part B: Applications
Volume **14**, Number **6**, June **2023**

ICIC International ©2023 ISSN 2185-2766
pp. 587–595

TOPIC-BOUND IMAGE CAPTION GENERATION: A MULTI-MODAL ENCODER-DECODER MODEL OF NEURAL NETWORKS WITH TRANSFORMER

Hidekazu Yanagimoto[1,*] and Kiyota Hashimoto[2,3]

[1]Graduate School of Informatics
Osaka Metropolitan University
1-1, Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
[*]Corresponding author: hidekazu@omu.ac.jp

[2]Faculty of Technology and Environment
Prince of Songkla University
80 Moo 1, Vichitsongkram Rd., Kathu, Phuket 83120, Thailand

[3]Faculty of Welfare and Information
Shunan University
843-4-2, Gakuendai, Shunan, Yamaguchi 745-8566, Japan
hash@g.shunan-u.ac.jp

Received November 2022; accepted January 2023

Abstract. *Caption generation is one of the multimodal learning tasks and deep learning contributes to the improvement of caption generation. When describing an image, humans can choose an object in focus as the topic and complete the description appropriately, which has not been achieved successfully by computer caption generation. Thus, it is important to develop a technique generating captions with the appropriate topic in terms of the object in focus. In this paper, we propose a topic-bound caption generation system with Transformer, which can generate captions including an object name in an image within the five top words of the caption. We make a caption generation corpus including images, captions, and focal points, which denote objects in images as topics of the captions. The results show the proposed method can generate topic-bound captions related to objects in an image and approximately 73.2% of all generated captions include the object names as a topic.*

Keywords: Deep learning, Natural language processing, Caption generation, Topic-bound caption, Text generation, Attention mechanism

1. **Introduction.** Various deep learning methods have been investigated to image and language processing [1, 2], and recently more attention is being paid to multi-modal processing such as caption generation of images and image generation hinted by keywords [3] by embedding multi-modal features into the same feature space. Based on this characteristic, many researchers pay attention to multi-modal learning [4] and tackle multi-modal learning tasks with deep learning. Caption generation [5, 6] is one of the multi-modal learning tasks, which generates a caption based on an input image. To realize caption generation, a caption must be generated based on image features extracted from an input image with an image recognition module. In other words, we have to make general-purpose features from the input image and it is related to representation learning [7] strongly. The caption generation system until now can generate captions of an image, but it is not controllable to choose which object in an image is focused on and generate a description accordingly, which humans can do almost unconsciously. Thus, it is an important next step for caption generation systems to be able to generate multiple captions with different choices of topics according to which object in an image is focused on. The caption topic

DOI: 10.24507/icicelb.14.06.587

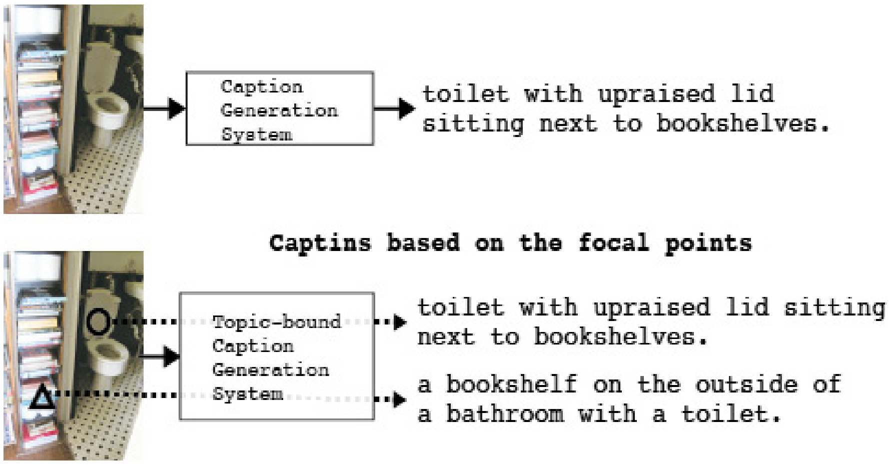

Caption Generation System
toilet with upraised lid sitting next to bookshelves.
Captins based on the focal points
Topic-bound Caption Generation System
toilet with upraised lid sitting next to bookshelves.
a bookshelf on the outside of a bathroom with a toilet.

FIGURE 1. Examples of topic-bound caption generation

depends on objects in an image and we present the objects as focal points in the image. Figure 1 shows the abstract of the caption generation task in this paper.

The first caption generation system [5] is implemented with an encoder-decoder architecture [8], which consists of VGG16 [9] as the encoder and a recurrent neural network as the decoder [8]. In this system, the encoder extracts features from an image and those features are embedded into the feature space. The decoder generates a caption based on the embedded features. A caption generation system with an attention mechanism [10] was proposed and every word in the caption is generated considering the image features [6]. The attention mechanism selects some of the image features depending on the hidden state in the decoder and utilizes the selected features to determine the next word in the caption. The latest caption generation system is OFA [11], which consists of ResNeXt [12] as the encoder and Transformer [13] as the decoder. All previous approaches can generate only a single caption from an image and it is almost arbitrary which object is chosen as the topic because the systems cannot consider multiple focal points in the input image. On the other hand, humans can determine focal points in the image by themselves and generate multiple captions based on the focal points. To simulate human behavior, the caption generation system in which the parameters are initialized with the focal points was proposed [14, 15]. In the caption generation system, image features are generated with VGG16 and GRU [16], which is a type of recurrent neural network, and generates a caption according to the focal points.

We propose a topic-bound caption generation system with a pre-trained object detection system and a Transformer-based language model system, which can generate a caption including object names in an input image based on focal points in the image. The proposed system replaces GRU in [16] with Transformer, which is the state-of-the-art neural network architecture in natural language processing. Transformer is more parallelizable than recurrent neural networks because it is able to relax order dependencies in word sequence by position encoding, and thus Transformer can enjoy a more parallelization effect with GPU than recurrent neural networks. Transformer does not include hidden states inside and we have to develop a mechanism to introduce the focal points into Transformer. Our proposal for this issue is that we utilize the focal points via the memory mask in Transformer and a context attention mechanism. Our contributions are as follows.

- We develop a topic-bound caption generation system with Transformer. Moreover, we introduce the context attention mechanism to the system and make the system generate topic-bound captions from the same image.

- We realize a Transformer-based caption generation system considering focal points in an input image: the system generates a caption including object names corresponding to the focal points within the 5 top words of the generated captions.

The organization of this paper is as follows. Section 2 describes our proposed models. Section 3 reports experiments: 3.1 for experimental conditions; 3.2 for experimental results and discussions. Section 4 gives the conclusions.

2. **Topic-Bound Caption Generation System with Transformer.** This section describes our topic-focused caption generation system with Transformer. The system consists of the pre-trained object recognition system, VGG16, and a Transformer-based language model.

2.1. **Image feature extraction with VGG16.** In our proposed system, VGG16 is employed for feature extraction from input images. Figure 2 shows the architecture of VGG16 and we use outputs from the 13th convolution layer as image features. Because VGG16 accepts images of 224 × 224 pixels, we transform the original images into the 224 × 224 images as below. First, an original image, I, is resized into a 256 × 256 image and cropped into a 224 × 224 image. After processing the image by VGG16, we obtain a 49 × 512 matrix as an image feature, F.

$$F = \mathrm{VGG16}(I) \tag{1}$$

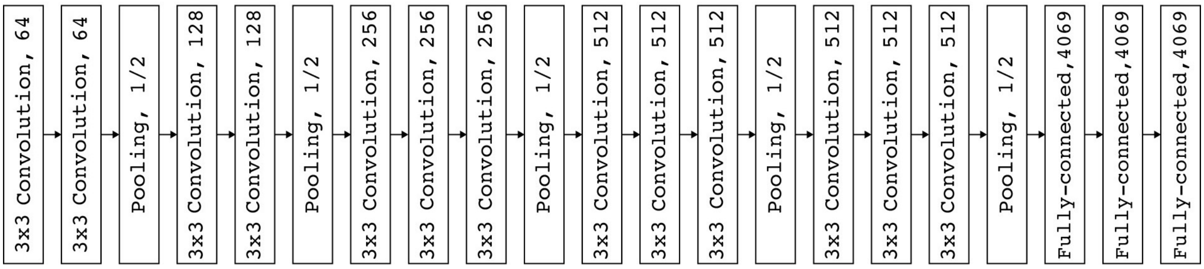


FIGURE 2. The architecture of VGG16

2.2. **Caption generation with Transformer.** The decoder part is a caption generation module with which a caption is generated based on the image features constructed by the encoder module. In this study the caption generation model is constructed with Transformer. Figure 3 shows the architecture of a decoder version of Transformer, which generates a sequence of words accepting external information in memory. The Transformer consists of two Multi-Head Attention modules and a feedforward neural network. Attention is defined with Equation (2).

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2}$$

The lowest Multi-Head Attention in Figure 3 accepts Q, K, and V, which are the same vector and is called self-attention. The other Multi-Head Attention in Figure 3 employs Q and K, which are the same vector but V is a different vector. So, the two Multi-Head Attentions work differently. A Multi-Head Attention is defined with Equation (3) and Equation (4).

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_n)W^O \tag{3}$$

$$\text{where } \mathrm{head}_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4}$$

First, in Equation (4), multiple attention is generated based on vectors, head_i, transformed with W_i^*. After then, in Equation (3), all attention is concatenated and transformed into the final feature vectors.

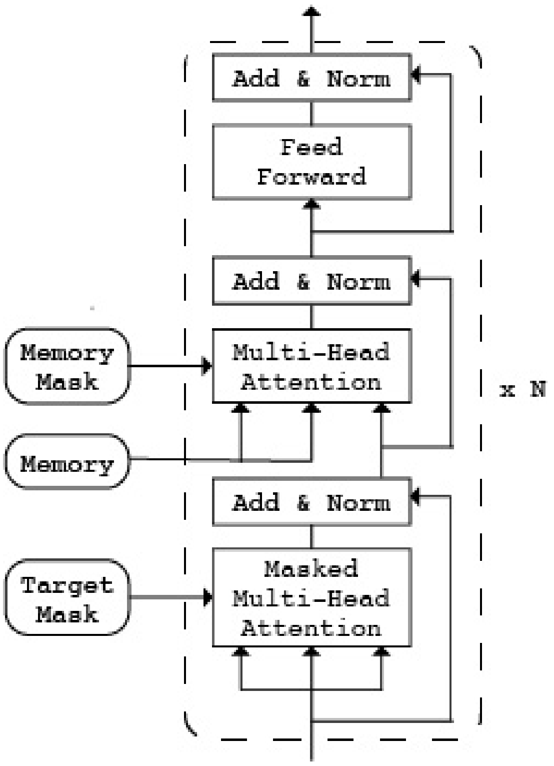


FIGURE 3. The architecture of Transformer (decoder version)

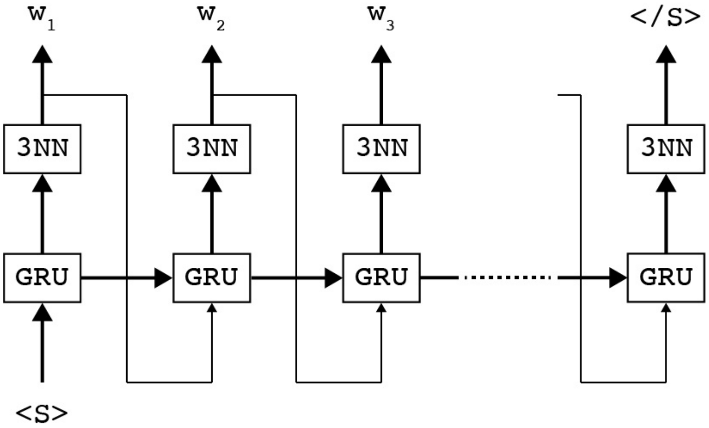


FIGURE 4. The architecture of GRU-based decoder

One of the motivations to adopt Transformer is parallelizability: the recurrent neural networks cannot work in parallel but the Transformer can process input data independently because there are order dependencies in the recurrent neural network. The GRU unit in Figure 4 must take over hidden states between GRUs. Additionally, Transformer employs position encoding to consider the order of the input data and can process input data in parallel. The position encoding is defined with Equation (5).

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin\left(\mathrm{pos}/10000^{\frac{2i}{d_{\mathrm{model}}}}\right) \quad (5)$$
$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos\left(\mathrm{pos}/10000^{\frac{2i}{d_{\mathrm{model}}}}\right)$$

The position encoding has different values according to their positions in the input data and is added to the input data.

The whole architecture of the caption generation model is shown in Figure 5. The image features, which are generated with VGG16 from an input image, are used as memory in Figure 3. To keep sound causality that each word choice is determined only by the previous words and not by consulting words that might follow, the target mask is employed in order to hide following words. The memory mask chooses the image features to generate a caption. To select the image features according to focal points, we employ the following memory mask.

$$\text{MemoryMask}_1 = [0, \ldots, 1, \ldots, 0] \qquad (6)$$

$$\text{MemoryMask}_i = [1, \ldots, 1] \ (i \neq 1)$$

MemoryMask_1 has 1 in focal points and 0 otherwise. The focal point denotes the position of a focusing object in the image. After applying VGG16, the image is resized into a $7 \times 7 \times 512$ tensor. Because the image features are flattened as a 49×512 matrix, the size of MemoryMask_1 is 49. So, the focal point is restricted from 0 to 48.

In Figure 5, we add context attention to the caption generation module. The context attention is similar to the attention mechanism in machine translation [17] and generates a context vector for every word in a caption, and the context vector is inputted in the

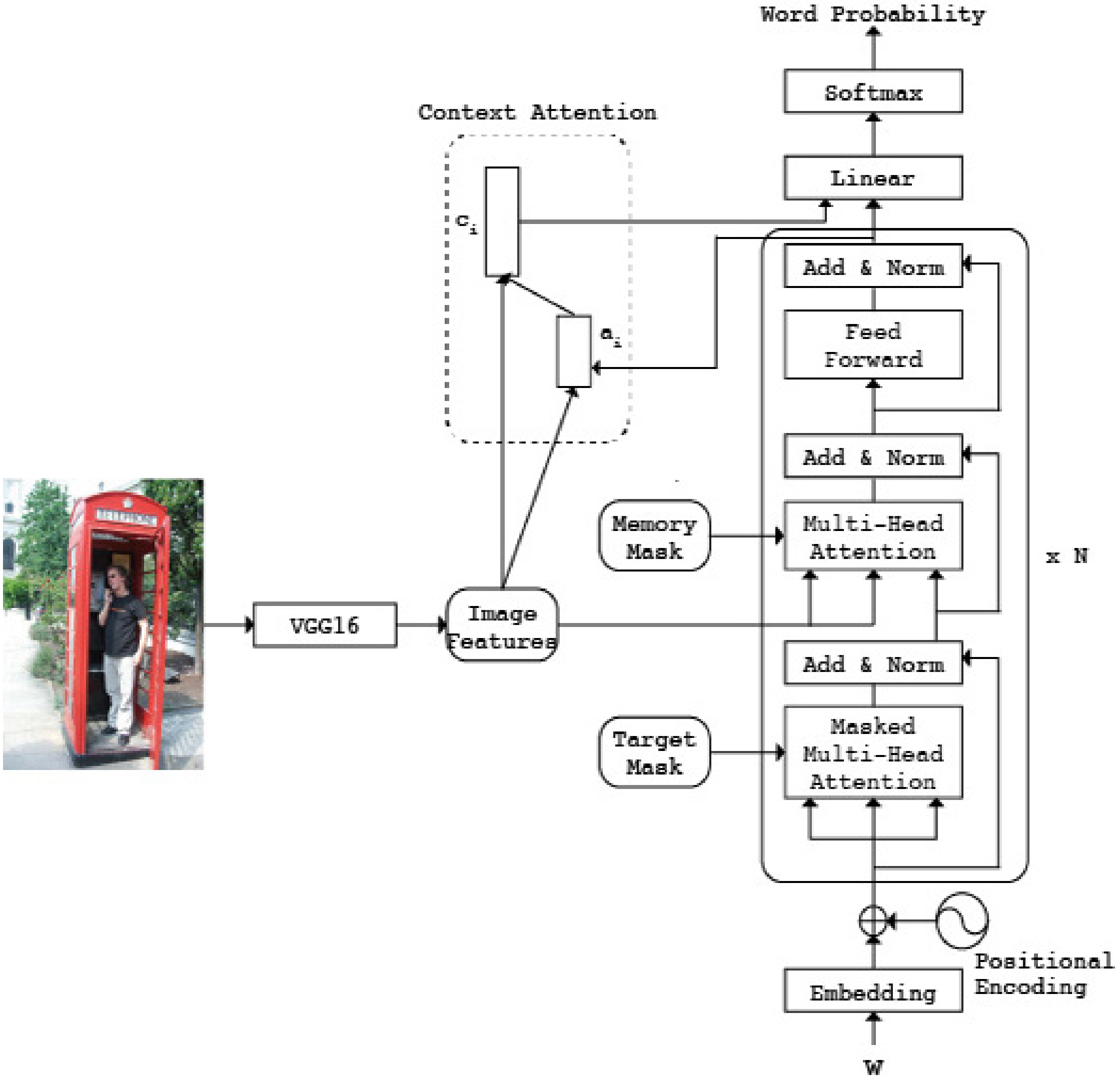


Figure 5. The architecture of the topic-bound caption generation system with Transformer

final fully-connected layer.

$$\mathbf{a}_i = F^{\mathrm{T}}\mathbf{h}_i \tag{7}$$
$$\mathbf{c}_i = F\mathbf{a}_i$$

The $\mathbf{h}_i$ denotes the output of the i-th hidden state in the Transformer-based decoder. The attention weight, $\mathbf{a}_i$, is calculated with the image features and the hidden state and the context vector is a weighted sum of the image features.

2.3. **Focal points in the image.** The focal points are determined as the position of a focusing object in an input image. The object that the focal point indicates is the topic of the caption of the image. VGG16 integrates local information of the image in each layer and constructs tensor-based image features. So, we can estimate correspondence between the focal point in the image and the image feature in the tensor. Figure 6 shows how to embed focal point information in tensor-based image features.

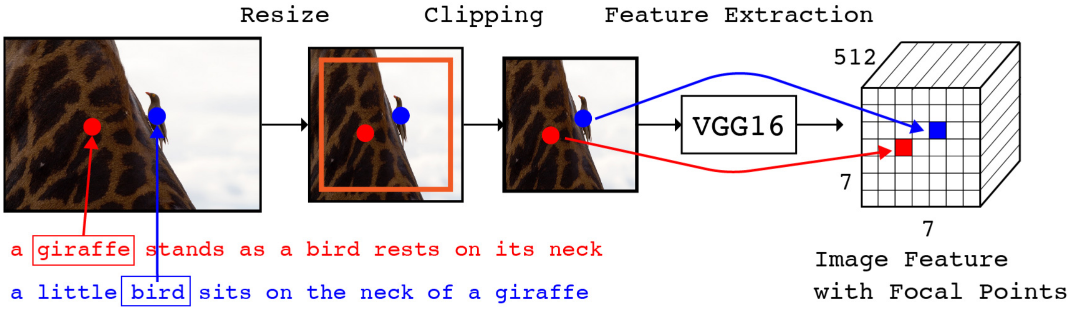

Resize
Clipping
Feature Extraction
VGG16
512
7
7
a giraffe stands as a bird rests on its neck
a little bird sits on the neck of a giraffe
Image Feature
with Focal Points

Figure 6. Focal point embedded image features

We regard the topic of the caption as a noun in the five top words of the caption. For example, we regard “giraffe” as a topic word for “a giraffe stands as a bird rests on its neck” and “bird” as a topic word for “a little bird sits on the neck of a giraffe”.

MS COCO dataset [18] contains about five captions for an image and detected object areas in the image. The focal point is the center of the detected object area in the image and is tied with the caption that includes the object name in the five top words of the caption. Using this focal point generation strategy [15], we construct image caption data with focal points automatically.

3. **Experiments.** We evaluate the topic-bound caption generation system with MS COCO dataset. In experiments, the system generates some different captions according to the focal points and we evaluate whether the system generates a caption including the topic word within the five top words of the caption.

3.1. **Experiments conditions.** MS COCO corpus is used for the evaluation experiments and consists of 82,783 training images and 40,504 validation images. Approximately five captions are prepared for an image. In the experiments, we use the validation data as test data because the test data in MS COCO corpus have no captions and we cannot evaluate the generated captions with the proposed system. In the experiment, 50,342 captions for 46,022 training images and 24,906 captions for 22,791 validation images are obtained, and the evaluation is made with how many of them chose the focal points that are used for the ground truth data. The focal point selection accuracy achieves 94.3% by a manual judgment of 1,629 images selected randomly. The hyper-parameters setting in the proposed method is shown in Table 1.

Table 1. Hyper-parameters settings

Parameters	Setting
Word embedding size	512
Hidden state size in Transformer	512
Stack size of Transformer	6
Vocabulary size	3,978
Minimum word occurrence frequency	3
Optimization algorithm	Adam
Learning rate	0.0001
Learning epochs	100

3.2. **Results and discussions.** First, we trained the proposed model with 50,342 training data.

Figure 7 shows a learning curve of the proposed method with training data including the focal points. The training error rapidly decreases during about 40 epochs and converges at about 0.1.

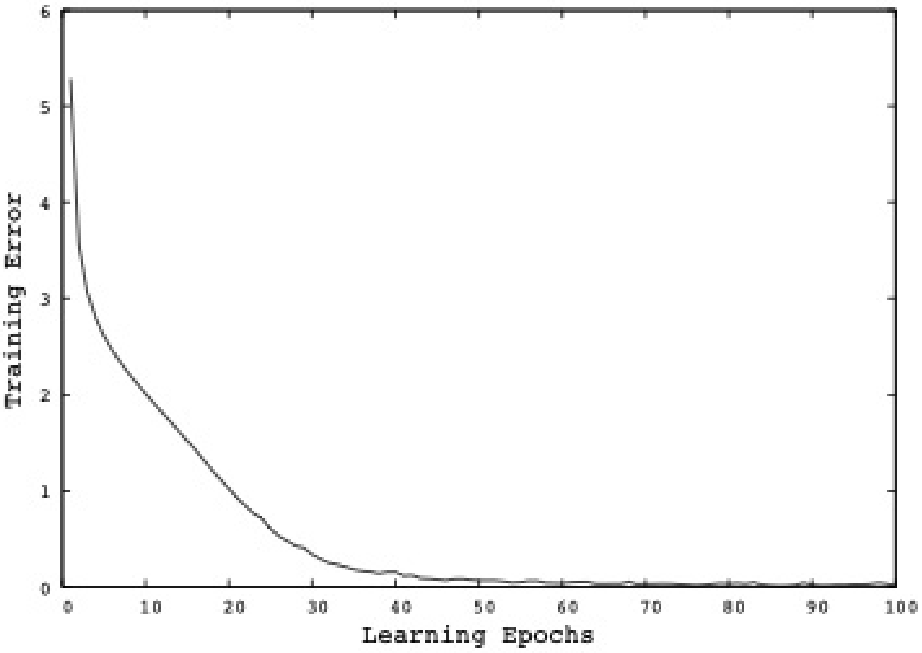
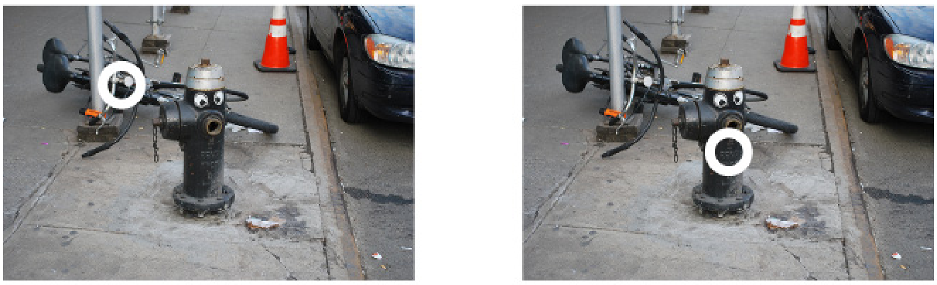


Figure 7. Learning curve of the proposed model training

Figure 8 shows generated captions by the proposed method according to the focal points. Object names in the image appear within the five top words in the caption. However, the caption is not correct grammatically and we need to train the proposed model with more datasets and more learning epochs.

We compare the proposed method with GRU-based caption generation [14]. The performance of the system is evaluated based on whether the object name, which the focal point denotes, is included in the five top words in the predicted caption. Table 2 shows the captions including the topic word with the proposed method and with a GRU-based method [14].

The proposed method can make more captions including objects pointed by the focal points than GRU-based method. It means that Transformer learns how to generate a caption considering the focal point well. It is not known yet how many stacks of Transformers are appropriate, but in our preliminary experiments, six stacked Transformers



several bicycles are parked on the pavement of a sidewalk.

a fire hydrant with a backpack on it.

FIGURE 8. Results of topic-bound caption genetaion

TABLE 2. Hyper-parameters settings

Method	Topic word inclusion rate
The proposed method	73.2%
GRU-based method [14]	44.9%

achieved enough in terms of function description, though theoretical discussion should be future work.

4. **Conclusions.** We proposed a topic-bound caption generation system with Transformer, which can generate captions according to focal points in an image. First, the proposed system can generate different captions for different focal points and the captions include object names pointed out by the focal points within the 5 top words of the generated captions. We compare the proposed method with a GRU-based caption generation system [14] and confirm that the proposed method can generate more favorite captions. Especially, in the proposed method approximately 73.2% of all generated captions include object names in an image within the 5 top words of the captions.

Future works are as follows. In the encoder module, VGG16 was employed in this study, mainly because of its popularity in studies of caption generation, but it naturally does not prove this choice as best, and comparisons among different image encoding methods are one of the future works. It is also a future work to investigate how deep Transformer should be stacked, not just from results but also from theoretical perspectives.

REFERENCES

[1] A. Candra, Wella and A. Wicaksana, Bidirectional encoder representations from Transformers for cyberbullying text detection in Indonesian social media, *International Journal of Innovative Computing, Information and Control*, vol.17, no.5, pp.1599-1615, 2021.

[2] Y. Chen, Y. Guo, H. Jiang, J. Ding and Z. Chen, Self-attention based Darknet named entity recognition with BERT methods, *International Journal of Innovative Computing, Information and Control*, vol.17, no.6, pp.1973-1988, 2021.

[3] R. Rombach, A. Blattmann, D. Lrenz, P. Esser and B. Ommer, High-resolution image synthesis with latent diffusion models, *Proc. of the IEEE/CVF Conferences on Computer Vision and Pattern Recognition*, pp.10684-10695, 2022.

[4] T. Baltrusaitis, C. Ahuja and L. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.41, no.2, pp.423-443, 2019.

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

[6] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Couville, R. Salakhutdinov, R. S. Zemel and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *Proc. of the 32nd International Conference on Machine Learning*, vol.37, pp.2048-2057, 2015.

[7] Y. Bengio, A. Couville and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.35, no.8, pp.1798-1828, 2013.

[8] I. Sutskever, O. Vinyals and Q. V. Le, Sequence to sequence learning with neural networks, *Proc. of the 27th International Conference on Neural Information Processing Systems*, vol.2, pp.3104-3112, 2014.

[9] K. Simonyan and A. Zisserman, Very deep networks for large-scale image recognition, *Proc. of International Conference on Learning Representations (ICLR2015)*, 2015.

[10] D. Bahdanau, K. H. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, *Proc. of International Conference on Learning Representations (ICLR2015)*, 2015.

[11] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou and H. Yang, Unifying architectures, tasks and modalities through a simple sequence-to-sequence learning framework, *arXiv.org*, arXiv: 2202.03052, 2022.

[12] S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He, Aggregated residual transformations for deep neural networks, *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] A. Vaswani, L. Jones, N. Shazeer, N. Parmar, J. Uszkoreit, A. N. Gomez and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing System*, vol.30, pp.6000-6010, 2017.

[14] H. Yanagimoto and M. Shozu, Multiple perspective caption generation with attention mechanism, *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, Kitakyushu, Japan, pp.110-115, DOI: 10.1109/IIAI-AAI50415.2020.00031, 2020.

[15] H. Yanagimoto and T. Imai, Multiple-perspective caption generation with initial attention weights, *Proc. of the 10th International Conference on Computer and Communications Management*, pp.19-23, 2022.

[16] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *Proc. of NIPS 2014 Workshop on Deep Learning*, 2014.

[17] T. Luong, H. Pham and C. D. Manning, Effective approaches to attention-based neural machine translation, *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp.1412-1421, 2015.

[18] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick and P. Dollar, Microsoft COCO: Common objects in context, in *Computer Vision – ECCV 2014. Lecture Notes in Computer Science*, D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars (eds.), Cham, Springer, DOI: 10.1007/978-3-319-10602-1_48, 2014.