Image Captioning in Near Future from Vehicle Camera Images and Motion Information

Yuki Mori[1,*], Tsubasa Hirakawa[1], Takayoshi Yamashita[1], Hironobu Fujiyoshi[1]

***Abstract*— Image captioning is a task to generate the sentence explaining an input image. In autonomous driving, image captioning is expected to be applied to provide linguistic explanations of autonomous driving control's decision-making because it can reduce the psychological burden on passengers and prevent accidents. The existing captioning methods have been limited to generate the caption for an input image and have not focused on generating captions for events in the near future. Regarding autonomous driving applications, it is important to generate captions for any events that will happen in the near future to prevent accidents and alert passengers. Therefore, in this paper, we propose a new task that generates the explanatory sentence of near future using images observed from past to present. To realize the near future caption generation in autonomous driving applications, we propose a near future image captioning method being suitable for in-vehicle camera images. Our experiments using the Berkeley Deep Drive eXplanation Dataset show that the proposed method can appropriately generate near future captions.**

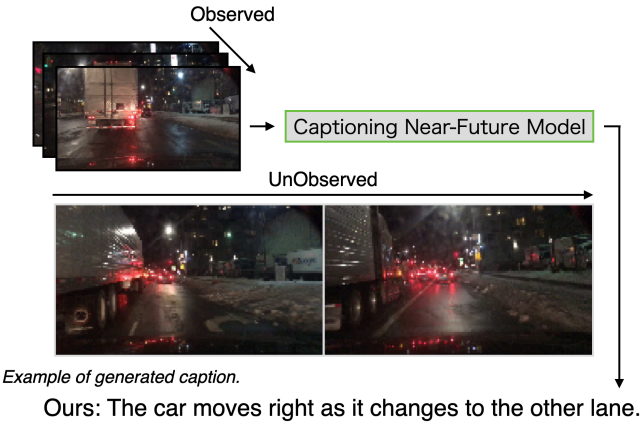

Fig. 1. The overview of near future image captioning. Our method inputs in-vehicle camera image sequences observed from past to present and generate captions explaining future events.

I. INTRODUCTION

Image captioning is one of computer vision task that generates a sentence explaining situation and event from an input image. Many methods using CNN and RNN have been proposed for image caption generation [1], [2], [3], [4], [5]. These methods employ an encoder-decoder model, which consists of a CNN that encodes the image into a feature vector and an RNN that decodes the encoded feature vector into a natural language caption. This allows to generate more natural captions corresponding to an input image.

Because of recent development of captioning methods, it has been proposed that image caption generation is applied to improve the explainability of autonomous vehicles [6] and to realize driving support systems for the purpose of accident prevention [7]. These existing image captioning methods generate sentences explaining events that have occurred in the past or present based on the input in-vehicle camera images. Meanwhile, for more practical autonomous driving applications, explaining events that will happen in the near future, such as motion of front vehicle or pedestrian after few seconds, improves the explainability of autonomous driving models and appropriately call a passenger's attention. For that reason, although it is necessary to develop an image captioning method explaining near future events, the conventional captioning methods does not focuses on such near future events.

In this paper, we propose a novel task: *near future image captioning*, which generate sentences explaining events that will occur in the near-future. To this end, we propose a method for generating near future captions from in-vehicle camera images. In addition to in-vehicle camera images, the proposed method introduces motion information of a vehicle and predicts future motions. The proposed method generates caption considering the predicted future motion and features encoded from input images, which can generate captions considering near future information. Our experiments using the Berkeley Deep Drive eXplanation (BDD-X) Dataset [6] show that the proposed method can appropriately generate near-future captions.

Our contributions are as follows:

- We propose a novel task: *near future image captioning*. The conventional captioning methods generate sentences describing events observed in-vehicle camera images. Meanwhile, the near future image captioning generates a sentence that explain events that will occur shortly from observed images.
- Captions for near future events can be used for improving the explainability of deep learning-based autonomous driving models and driver assistance systems for accident prevention and alerting passengers.

II. RELATED WORK

Image captioning is the task to generate a suitable description for a given image. Based on the problem, captioning methods summarize the input image or explains the basis for the decision-making of a network output. Image captioning has been studied before deep learning attracted attention [8],

* Corresponding author. yukiri@mprg.cs.chubu.ac.jp
[1] Authors are with Chubu University, Kasugai, 487-8501, Japan.

[9]. Since the advent of deep learning, multiple captioning methods utilizing CNN and RNN have been proposed [10], [11], [12], [13], [14]. Show and Tell [10] is a method for generating image captions that consists of a CNN as the encoder and an RNN as the decoder. In the encoder, image features are extracted from the input image. The image features are then input to the decoder, which outputs the appropriate words. The output word becomes the input for the next time, and outputs the next suitable word. The image features and language features are learned to be obtained in the same embedding space. LSTM is used for the RNN of the decoder.

In image caption generation using LSTM, a longer input sentence disturbs efficient propagation of information, which causes the accuracy deterioration. To overcome this problem, Show, Attend and Tell [11] proposes an attention mechanism. The proposed attention mechanism suppress the decrease in accuracy and achieved accurate caption generation.

While the above methods are intended for general images, the image captioning methods for in-vehicle camera images have been also proposed [6], [7].

In the case of in-vehicle camera images, it is possible to prove the basis for decisions using natural language for automatic driving control, and to alert passengers of the surrounding situation.

Kim et al.[6] have proposed a method for generating captions from in-vehicle camera images. This method is constructed from a Vehicle Controller model that learns and infers the movement information of the own vehicle from the input image, and a Textual Explanation Generator model that explains the behavior of the own vehicle. The attention obtained by the Vehicle Controller is used by the Textual Explanation Generator to enable caption generation based on the vehicle's behavior.

Mori et al.[7] proposed Attention Neural Baby Talk. This method uses an object detector to detect risk factors in the surrounding environment of a vehicle, incorporates the detection results into image caption generation, and generates captions to communicate to the passengers. In this method, risk factors are selected by a rule base and an object detector, and attention masks are applied to the input features to generate image captions to alert the driver about specific risk factors.

The conventional image caption generation method for in-vehicle camera images has been developed so that the image caption generation for general images can be used for in-vehicle camera images. Therefore, it is possible to provide a linguistic explanation of the basis for the decision at this time. On the other hand, linguistic explanations for accident prevention and risk factors, such as alerting passengers, need to focus future events, not current ones. The problem with conventional methods is that they do not generate image captions for such near-future events.

III. PROPOSED METHOD

In this study, we propose a new task: *near future image captioning*. We propose a caption generation model that is suitable for this task. In this chapter, we first explain the new task of near future caption generation, and then describe the network structure and training method of the proposed method.

A. Near-Future Image Captioning

To develop image captioning with in-vehicle camera images for preventing traffic accidents and alert passengers, we need to explain the situations in the near future, such as the movement of front vehicle and pedestrians after few seconds. Because the conventional captioning methods generate sentences for given images, these methods cannot consider the information in the near-future, as shown in Fig. 1. In the proposed task, we input multiple images observed from past to present into a near future captioning model and generate captions considering the future movements. Here, note that we do not use unobserved images corresponding to the near future. A captioning model captures notable areas for the near-future events from observed images and generates caption.

B. Near future captioning model

Figure 2 shows the model structure for generating captions in near future. First, the model extracts feature vectors from multiple images. Then, the extracted features and a sensor data representing the vehicle motion information are incorporated and input to an encoder. Incorporating sensor data enables to consider vehicle information that cannot be captured from images. The action regressor predicts vehicle motion from intermediate features obtained from the encoder. From the intermediate feature and predicted future vehicle motion, caption decoder generates caption in the near-future. Hereafter, we introduce the details of the proposed model.

1) Feature extraction network: The feature extraction network consists of five convolutional layers. We use ReLU function as the activation function for each layer. We input images into the feature extraction network and the extract feature vectors. The feature vectors of each time are individually extracted. The extracted feature is then input to the encoder as observation information at each time.

Note that the feature extraction network is pre-trained using the BDD-X dataset. Because the BDD-X dataset includes vehicle information such as the acceleration and angle of the vehicle, we train the feature extraction network to estimate the vehicle information in advance. By pre-training, it is possible to extract features suitable for caption generation of in-vehicle camera images. We use the first five convolutional layers excepting for the last output layer in the pre-trained network as the feature extraction network.

2) Encoder: The encoder consists of two LSTM layers. We input the feature vectors of each time extracted by the feature extraction network. Since we input the n feature vectors sequentially, the obtained intermediate representation considers the time series changes of input feature vectors. The number of units of each LSTM layer is 1,024.

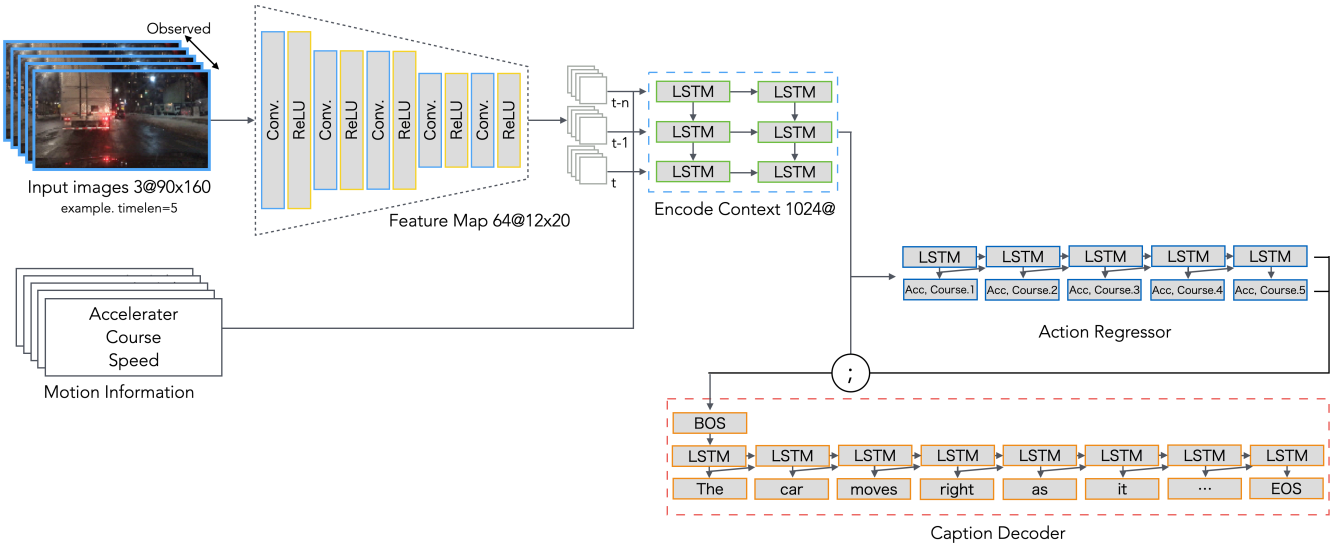

Fig. 2. The proposed network architecture of near future image captioning

3) Motion information: In order to capture changes of the camera image sequence in time series, we input motion information acquired from sensors to the encoder in addition to the input images. We use velocity, acceleration, and steering angle as input motion information. By using both velocity and acceleration, the feature takes into account the degree of acceleration and deceleration at any given speed. Also, by using the steering angle, it is possible to consider whether the vehicle is going straight or turning a curve.

4) Action regressor: Action regressor estimates the motion information of the own vehicle in the near future. This network consists of an LSTM and three fully connected layers. We input the intermediate representations acquired by the encoder, and the velocity and steering angle of the car are predicted as the future motion information. The number of units in each layer of this network is 1,164 in the first layer, 100 in the second layer, 50 in the third layer, and 10 in the fourth layer. Action regressor outputs the velocity and steering angle for m frames. The output of the action regressor is combined with the intermediate representation acquired by the encoder and fed into the decoder.

5) Decoder: The decoder consists of a single layer of LSTM and a fully connected layer. The 1,024 dimensional intermediate representation acquired by the encoder and the 10 dimensional output of the action regressor are input to the decoder, and then the decoder outputs a word sequence that explains events in the near future. The number of input dimensions is 1,034, and the number of LSTM units is 1,024. The fully connected layer outputs the occurrence probabilities of each word. The number of units is M, which is the size of a dictionary.

C. Training

The proposed method consists of a feature extraction network that acquires feature vectors from images, an encoder that converts image feature vectors into intermediate representations, an action regressor that estimates the near future motion information of the own vehicle, and a decoder that generates near future captions. The feature extractor network and the encoder-decoder part are trained separately.

We first train only the feature extraction network using the BDD-X dataset. As the pre-training task, we train the network to estimate vehicle speed and steering angle from an input image. We use a mean squared error as the loss function. After the pre-training, we use only the first five convolutional layers for the feature extraction network.

The action regressor and the encoder-decoder parts are trained in an end-to-end manner. As loss functions for each module, we use a mean squared error for the action regressor and a cross entropy loss for the encoder-decoder. Here, we denote a_t and c_t as speed and steering angle of the own vehicle, respectively. Let x_k and h_k be the outputs of the encoder LSTM and intermediate layer of the decoder LSTM, the loss function L is defined as follows:

$$L = \sum_t ((a_t - a'_t)^2 + (c_t - c'_t)^2) + \sum_k \log p(y_k \mid y_{k-1}, h_k, x_k), \quad (1)$$

IV. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed method. We use the BDD-X dataset to evaluate the performance of the proposed method for generating near future captions.

A. Experiment Summary

We use the Berkleley Deep Drive eXplanation (BDD-X) dataset [6], which consists of 6,984 in-vehicle camera images. The BDD-X dataset includes annotations with respect to vehicle motions such as velocity, acceleration, angle of travel, and 26,228 control events of the car. The control

TABLE I
FREQUENT WORDS IN THE DESCRIPTION SECTION

Word	Count
stop	6879
slow	6122
forward	4322
drive	3994
move	3273

TABLE II
FREQUENT WORDS IN THE EXPLANATION SECTION

Word	Count
traffic	7486
light	6116
red	3979
move	3915
clear	3660

event indicates the time of vehicle control event and the caption annotation of the car's behavioral reason explanation. The caption annotation consists of a description part that verbalizes the behavior of the vehicle and an explanation part that verbalizes the reason for the behavior. Typical events are acceleration/deceleration, turning left/right, changing lanes, merging, and retreating. In addition, the reasons for each event are that the front is vacant, high-speed merging, and parallel parking. The number of dictionary words in the dataset is 1,290. The five most frequently occurring words in description and explanation parts are shown in Tables I and II, respectively. The description part contains many words related to acceleration and deceleration, such as stop and forward. The explanation part contains many words related to the situation in front, such as the color of traffic lights, move and clear.

The video frame rate is 30 fps and the average time of the events is 7.26 seconds for the whole data. In our experiment, we adjust to 1 fps to reduce the computational costs. We use 4,356 training samples with 14,933 sentences, and 536 evaluation samples with 1742 sentences.

In this experiment, the number of training epoch is 30, and the batch size is 50. The image size is 160×90 pixels, and the input frame n is 5. The number of output units of the decoder M is 1,290. We initialize the network parameters by Xavier initialize algorithm. We Adam as an optimizer for each network modules. As evaluation metrics, we use BLEU [15], METEOR [16], and CIDEr [17].

B. Definition of caption generation time

The BDD-X dataset does not contain caption annotations in the near future. Therefore, we use the existing caption annotations as the captions for the near future by defining the temporal intervals of caption generation as follows.

- Current: Input the entire event occurrence interval and generate caption
- Near future: Input the first half of the event occurrence interval and generate the caption for the second half.
- Future: Generate captions during the event interval using the information before the event occurred.

The 'current' uses the image during the event as input to generate the caption. When the caption is generated, the event is over. This is equivalent to conventional caption generation. The 'near future' produces captions for events that will occur in the near future a few seconds later. The 'future' generate captions during the event interval using the information before the event occurred. In the case of 'near future', we expect to be able to generate captions about whether the car will stop in the future, along with the rationale for the initial decrease in speed during events such as braking to a stop.

TABLE III
ACCURACY COMPARISON BY CAPTION GENERATION TIME

Time	BLEU@4	METEOR	CIDEr
Current	15.97	28.20	74.96
Near-Future	**17.02**	**29.26**	**83.73**
Future	11.97	27.70	47.11

TABLE IV
EVALUATION OF THE USEFULNESS OF MOTION INFORMATION AND ACTION REGRESSOR

Model	Time	BLEU@4	METEOR	CIDEr
None	Current	12.10	26.92	45.47
	NearFuture	12.83	26.56	49.04
Action reg.	Current	13.75	27.91	59.35
	NearFuture	13.17	26.69	51.96
Motion info.	Current	16.16	28.67	75.61
	NearFuture	16.74	28.77	77.33
Motion info. and action reg.	Current	15.97	28.20	74.96
	NearFuture	**17.02**	**29.26**	**83.73**

C. Evaluation over different caption generation time

In this experiment, we compare the accuracy at each caption generation time defined in the previous section. Furthermore, the usefulness of the motion information of the own vehicle and the action regressor introduced by the proposed method is confirmed at each generation time. Table III shows the accuracy of caption generation of each caption generation time. With respect to the BLEU@4, 'near future' with the first half of the event occurrence interval as input has the best accuracy 17.02, indicating that the caption generation is suitable for 'near future'. In addition, the evaluation indices of METEOR and CIDEr are 29.26 and 83.73, respectively, indicating that the accuracy has been improved compared to the 'current' method that uses the entire event as input. On the other hand, the accuracy of 'future', which is input before the event occurrence section, is significantly reduced to 11.97 in the evaluation index of BLEU@4. The reason would be that it takes a long time for the event to occur and the changes in the image and motion information do not match the event. This shows that the prediction of the future for a long time ahead is more challenging task, achieving accurate captioning for 'future' is one of our future work.

D. Evaluation of motion information and action regressor

Table IV shows the comparison of the accuracy of caption generation with and without motion information and the action regressor. We can see that the accuracy in the near future is improved to 16.74 in BLEU@4 by adding motion information compared to the case where motion information is not used. In particular, results of CIDEr show a significant improvement from 49.04 to 77.33 in generating captions for the near future. When the action regressor is added, the

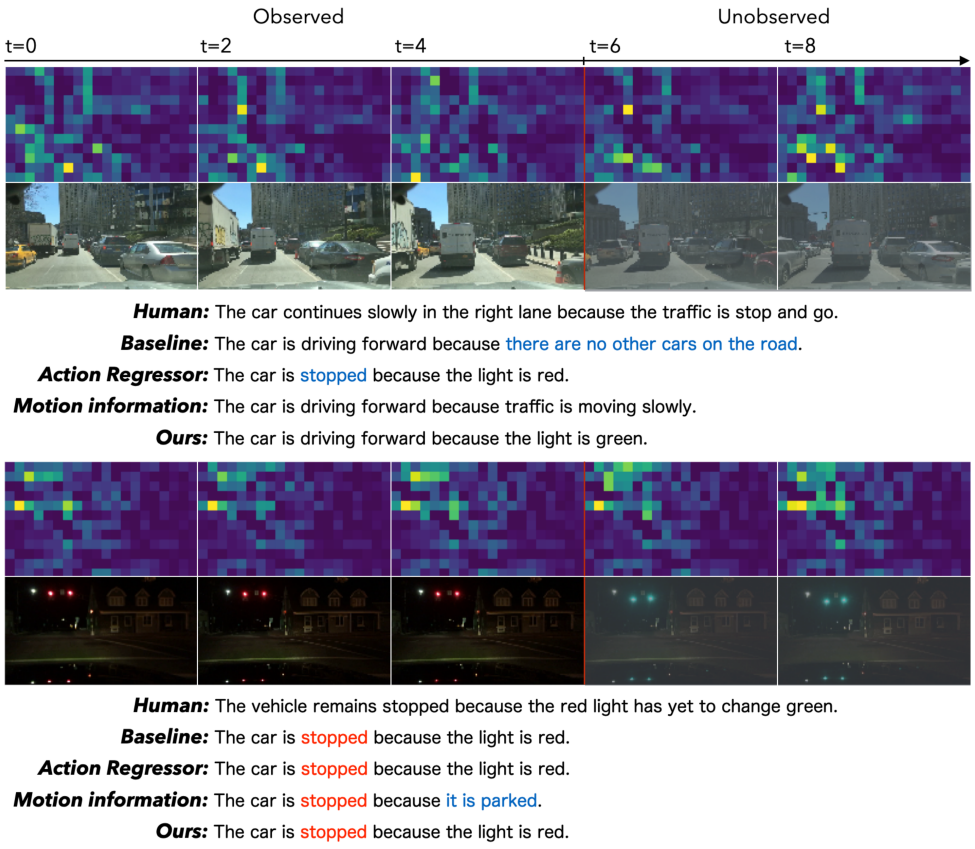

Fig. 3. Feature visualization and generated captions in turning scenes.

TABLE V
ACCURACY OF ACTION REGRESSOR'S MOTION PREDICTIONS

Model	Acceleration (m/s^2)	Course (degree)
Action regressor	6.76	9.22
Motion information and action regressor	5.05	4.73

accuracy of each evaluation index is improved as well, but it is lower than that of the motion information alone. It is considered that the caption generation accuracy is reduced when the speed and steering angle of the action regressor are not predicted correctly. On the other hand, when both motion information and action regressor are added, the accuracy of the BLEU and METEOR evaluation indices is slightly lower than when only motion information is added, but the accuracy of the CIDEr evaluation index is 78.94, which is better than when only motion information is added. CIDEr is an index that can consider the evaluation of ambiguous expressions of the same word. It is thought that the expressive power of caption generation is improved by introducing motion information and the action regressor.

Table V shows the accuracy of action regressor. Table V shows that the accuracy of the Action Regressor is better when motion information is used as input than when motion information is not used as input. This suggests that the input of motion information helps to improve the accuracy of Action Regressor and the expressiveness of caption generation.

E. Qualitative evaluation of generated captions

We compare the captions generated by each comparison method and visualize input features. We set the near future as the generation time, and the results of the visualization are shown in Figures 3 and 4. The gray images are the images in the range of caption generation, which are not used as input in our proposed method.

Figure 3 shows the results of the scene in which vehicle turns right. In the results of the top scene, while the captions without motion information describes "driving forward" or "stopped", the caption with motion information successfully describes "turning right." In addition, the image features highly respond to the area around the traffic lights and the front vehicle. In the bottom result, the bose captions of motion information and proposed method describes about turn to the right. In particular, the proposed method is able to generate the expression of slowing down.

Figure 4 shows the results of the scenes in which vehicle waits for a traffic light. From the top results, image features are responsive to the vehicle in front. In the first input frame, the brake lights of the car in front are lighted, but not in the last frame. Without using motion information, incorrect captions are generated. Meanwhile, when motion information is used as input, captions describing acceleration

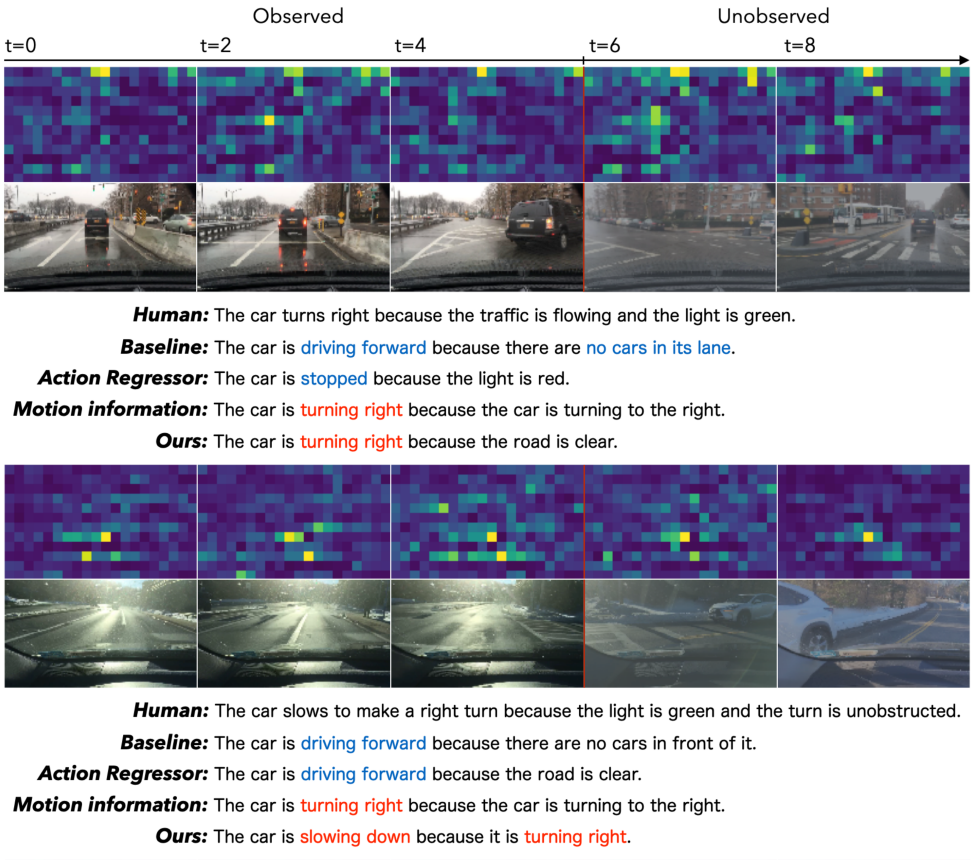

Fig. 4. Feature visualization and generated captions in waiting scenes.

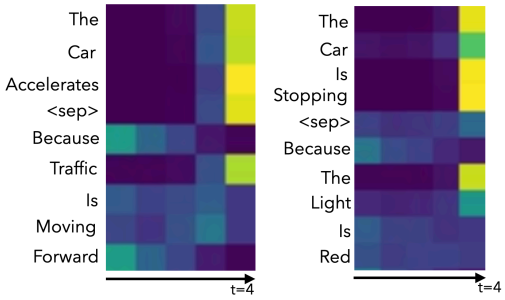

Fig. 5. Example of Attention visualization for language generation

are generated. In the bottom results, the image features respond to the vicinity of traffic lights. In this scene, the expression "stopped" has been generated. However, in the unobserved part that should be predicted, the traffic light is green. Thus, even in the near future, it is difficult to predict when a traffic light will turn from red to green.

Consequently, it was found that captions suitable for events that occur in the near future can be generated by using images and motion information. As shown in these results, we can see that the caption is close to the caption created by a human as annotation when the motion information of the vehicle is used as input

F. Attention visualization for caption generation

We visualize the attention of each generated words over time. The attention is calculated by the soft-attention mechanism [11]. Figure 5 shows the attentions. We can see that the first half of the caption focuses on the last input time, while the second half of the caption focuses on all features at all times. Since the proposed method uses motion information, we can assume that the first half of the caption focuses on motion information, while the second half focuses on image features.

V. CONCLUSIONS

In this paper, we propose a new task: near future image captioning. We also proposed a model being suitable for generating image captions for near future events for in-vehicle camera images. Through evaluation experiments, we show that the proposed method can generate near future image captions. We also confirmed the effectiveness of the use of vehicle information such as vehicle acceleration and speed as input for image caption generation for in-vehicle cameras. Our future work includes applying near future image captioning for the other images except for in-vehicle camera images. We will also investigate the elements necessary for caption generation that considers the distant future.

REFERENCES

[1] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019, pp. 4125–4134.

[2] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019, pp. 10 685–10 694.

[3] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019, pp. 6241–6250.

[4] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 2019, pp. 8957–8964.

[5] F. Sammani and M. Elsayed, "Look and modify: Modification networks for image captioning," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019, p. 75.

[6] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *The European Conference on Computer Vision (ECCV)*, 2018.

[7] Y. Mori, H. Fukui, T. Hirakawa, J. Nishiyama, T. Yamashita, and H. Fujiyoshi, "Attention neural baby talk: Captioning of risk factors while driving," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 4317–4322.

[8] Y. Ushiku, T. Harada, and Y. Kuniyoshi, "Efficient image annotation for automatic sentence generation," in *ACM International Conference on Multimedia*, 2012, pp. 549—558.

[9] Y. Ushiku, M. Yamaguchi, Y. Mukuta, and T. Harada, "Common subspace for model and similarity: Phrase learning for caption generation from images," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2668–2676.

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

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, vol. 37, 2015, pp. 2048–2057.

[12] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 375–383.

[13] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7219–7228.

[14] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *IEEEConference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8307–8316.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311—318.

[16] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014, pp. 376–380.

[17] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.