Attention Neural Baby Talk: Captioning of Risk Factors while Driving

Yuki Mori[1,*], Hiroshi Fukui[1], Tsubasa Hirakawa[1],
Jo Nishiyama[2], Takayoshi Yamashita[1], and Hironobu Fujiyoshi[1]

***Abstract*— Driving has various risk factors, including the possibility of traffic accidents involving pedestrians and/or oncoming vehicles. A driver assistance system that can prevent traffic accidents must be able to get the driver's attention to enable better safety. A practical solution for attention attraction should involve caption generation from in-vehicle images. Although a number of approaches for caption generation with deep neural networks have been proposed, they are inadequate for the specific risk factors while driving. The reason is that conventional captioning methods focus on not these factors but the entirety of an image. To tackle this problem, we first created a dataset to attract attention, one that considers risk factors during driving. Furthermore, we propose an image captioning method for the assistance system. Our method is based on neural baby talk and introduces an attention mask focusing on risk factors in an image. The mask enables our model to generate captions on each factor. Experimental results with our created dataset show that our method can generate captions for ideal attention attraction.**

I. INTRODUCTION

Sufficient attention attraction for passengers in a vehicle is required to achieve a driver assistance system for preventing traffic accidents. Multiple risk factors occur during actual driving, e.g., those involving pedestrians and oncoming vehicles. Image captioning could be an effective way to inform the driver of such risk factors.

Such captioning generates descriptions of an image to depict the environment, situation, or existing objects in the image. After the development of convolutional neural networks (CNNs) [1] and recurrent neural networks (RNNs) [2], a captioning method combining the two has been widely studied [3], [4], [5], [6].

To train captioning models, we need a dataset that includes manually annotated correct captions for each image. The construction of such a captioning dataset involves a higher cost, and ensuring the quality of annotated captions is difficult. The reason is that the difference in the annotation quality is caused by the characteristics of each annotator. Moreover, most conventional image captioning models generate a caption for each image, e.g., Fig. 1. This is inadequate in case that several objects should be paid attention exist in an image.

This paper presents a solution to the aforementioned problems, a driving assistance system using image captioning. Specifically, two approaches are proposed. The first is a suitable dataset automatically created for attention attraction, one which selects risk factors in an image using object detection and rule-based attribute extraction. The second is an attention mask to neural baby talk (NBT) [6], which is an image captioning method. This enables us to generate multiple captions with respect to each risk factor in an image.

Fig. 1. Example of caption generated by neural baby talk [6].

Our contributions are as follows.

- We propose a method to create a captioning dataset suitable for driving using rule-based annotations. Specifically, our method automatically annotates ground truth for each image using rule-based attribute extraction. It enables us to create a dataset at lower cost.
- We propose a system to get the driver's attention focusing on risk factors during driving. Our method can generate multiple captions for an image, solving the problems of conventional approaches.

II. RELATED WORK

A. Show and Tell

A typical deep learning-based image captioning method was proposed by Vinyals et al. [3]. It uses a long short-term memory (LSTM) [7]. This method consists of these modules: a CNN module to extract a feature vector x_{-1} from an image, a module to convert words in a sentence into a feature vector W_e, and a module to compute the appearance probability of the next word p_{t+1} by inputting feature vector x_t into the LSTM. Here, let I be an image, S_0 be a symbol to start captioning, and $S = \{S_1, S_2, ..., S_{N-1}\}$ be an output result from LSTM at each step t. The feature extraction from an image is formulated as

$$x_{-1} = P^{CNN}(I), \tag{1}$$

*Corresponding author yukiri@mprg.cs.chubu.ac.jp
[1] Authors are with the College of Engineering, Chubu University, Kasugai, 487-8501, Japan
[2] Jo Nishiyama. Researcher with Nissan Motors, Yokohama, 220-8686, Japan

and the computations of feature vector x_t and appearance probability p_{t+1} are defined by

$$x_t = W_e S_t, \quad t \in \{0, ..., N-1\} \tag{2}$$
$$p_{t+1} = P^{LSTM}(x_t), \quad t \in \{0, ..., N-1\}. \tag{3}$$

B. Show, Attend and Tell

In image captioning with RNN or LSTM, the longer the handled sequence information is, the lower the captioning accuracy becomes. A captioning method that introduces an attention mechanism has been proposed to overcome this problem [4]. In the attention mechanism, the features extracted from a network are weighted to select important features, achieving higher accuracy. The attention mechanism has two approaches, soft and hard attentions. The former uses a weighted average of multiple vectors, and the latter selects one element from several.

C. Adaptive Attention

Captioning methods with an attention mechanism have a problem where features extracted from images may affect words that might not require image features, e.g., prepositions or conjunctions. To overcome this problem, Lu et al. [5] proposed an attention mechanism that adaptively decides whether or not the model uses image features to generate each word. The adaptive attention mechanism computes image features $V = \{v_1, v_2, ..., v_k\}$, $v_i \in \mathcal{R}^d$ using a CNN, which is divided into k grids. Then, it computes the weight for the attention mechanism $\alpha \in \mathcal{R}^k$ using the image feature V and hidden state of an LSTM h_t:

$$z_t = w_h^T \tanh(W_v V + (W_g h_t)\mathbb{1}^T) \tag{4}$$
$$\alpha = \mathrm{softmax}(z_t). \tag{5}$$

The feature vector c_t applying the weighted average by α is calculated using

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{ti}. \tag{6}$$

To decide if image features are used to generate a caption, we use a visual sentinel vector s_t. Here, let h_{t-1} and m_t be a hidden state and a cell state of an LSTM, respectively, and s_t is formulated using

$$g_t = \sigma(W_x x_t + W_h h_{t-1}) \tag{7}$$
$$s_t = g_t \odot \tanh(m_t). \tag{8}$$

s_t is calculated by extending the LSTM. Using the point-wise product of m_t and g_t, we can decide if we should consider image features.

The feature vector considering visual sentinel $\hat{c}_t$ is calculated using

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t, \tag{9}$$

where $\beta_t \in [0, 1]$ is a gate that decides the use of image features at time t. If $\beta_t = 0$, s_t is used for taking the weighted average, which is defined as

$$\hat{\alpha}_t = \mathrm{softmax}([z_t; w_h^T \tanh(W_s s_t + W_g h_t)]). \tag{10}$$

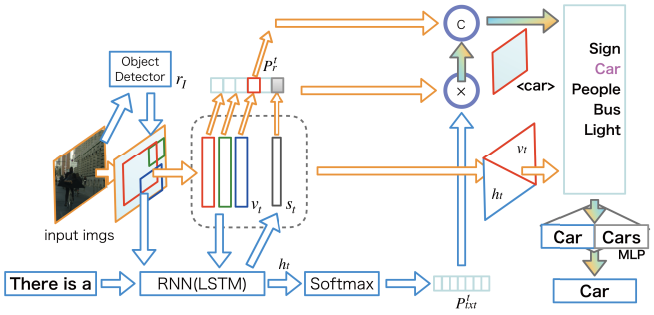

Fig. 2. Network architecture of NBT.

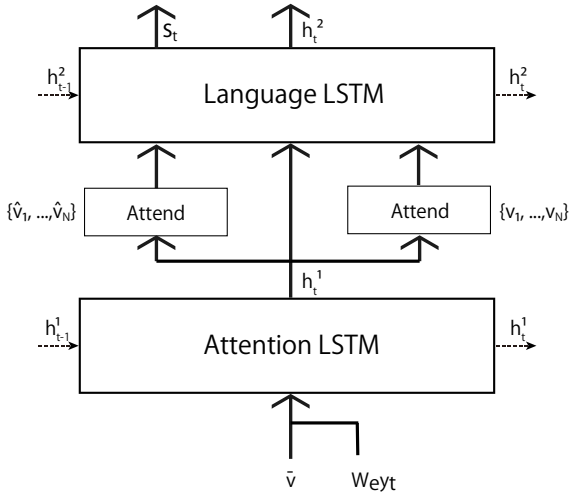

Fig. 3. Language model of NBT.

The appearance probability introducing the adaptive attention mechanism can be formulated using Eq. (9):

$$p_t = \mathrm{softmax}(W_p(\hat{c}_t + h_t)) \tag{11}$$

D. Neural Baby Talk

The adaptive attention mechanism achieves higher captioning accuracy by deciding the use of image features. Additionally, NBT [6] utilizes object detection. The network architecture of NBT is shown in Fig. 2. NBT detects the objects using the region proposal network (RPN) and RoI pooling. The features and labels of each object region are then used to generate captions.

NBT has a module to introduce the features and labels of the detected object region, which is achieved by estimating the labels of the object candidate region and generative probability y_{txt} of a language model. The probability distribution for captions of each object candidate region P_{rI}^t are calculated using pointer networks [8]. A pointer for input element u_i^t is defined by

$$u_i^t = w_h^T \tanh(W_v v_t + (W_g h_t)\mathbb{1}^T), \tag{12}$$

and P_{rI}^t is calculated using

$$P_{rI}^t = \mathrm{softmax}(u_i^t). \tag{13}$$

The probability to adopt the output of the language model y_{txt} is defined by

$$p(y_t^{txt}|y_{1:t-1}) = p(y_t^{txt}|\tilde{r}, y_{1:t-1})p(\tilde{r}|y_{1:t-1}). \quad (14)$$

We apply y_{txt} to adaptive attention. y_{txt} is calculated using visual sentinel vector s_t as Eqs. (8) and (8). Probability P_r^t to select the object candidate region is defined by

$$P_r^t = \mathrm{softmax}([u^t; w_h^T \tanh(W_s s_t + W_g h_t)]). \quad (15)$$

The last element of P_r^t is visual sentinel vector $\tilde{r}$.

Then, we apply the visual sentinel vector $tilder$ to $p(\tilde{r}|y_{1:t-1})$. The conditional probability of the language model output y_{txt} is defined by

$$P_{txt}^t = \mathrm{softmax}(W_q h_t). \quad (16)$$

By applying Eqs. (15) and (16) to Eq. (14), we can obtain the conditional probability of the language model output y_{txt} considering the visual sentinel. NBT generates a caption among P_r^t and P_{txt}^t by selecting a higher probability.

Moreover, unlike conventional attention models, NBT utilizes an attention model consisting of two LSTM layers [9]. Therefore, we can apply the attention mechanism toward each object candidate region instead of toward each grid. Figure 3 shows the language model of NBT. Here, let $V = \{v_1, v_2, \ldots, v_k\}, v_i \in \mathcal{R}^d$ be features of each object candidate region, and let $\hat{V} = \{\hat{v_1}, \hat{v_2}, \ldots, \hat{v_k}\}, \hat{v_i} \in \mathcal{R}^d$ be image features extracted by CNN divided into k grids. The weight of the attention mechanism can be calculated by Eq. (5).

In that way, NBT can generate accurate captions by introducing object detection.

III. PROPOSED METHOD

In this paper, we propose the following two approaches to achieve image captioning suitable enough to get the driver ' s attention, which is based on NBT. We first explain a rule-based automatic annotation method to create a dataset for image captioning of the attention attraction during driving. We then introduce an image captioning method by describing an attention mask to the NBT model. Our methods could solve the problems of the conventional captioning method and could generate captions focusing on risk factors.

A. Building a dataset to get a driver ' s attention

Existing datasets for image captioning are created manually. This causes several problems, such as high cost and difficulty ensuring the quality of annotated captions because of the high variation in characteristics between annotators. Furthermore, the existing datasets are not suitable for driving, and a new dataset needs to be built for ideal attention attraction. We propose an automatic annotation method to reduce the annotation cost.

Figure 4 shows an overview of our automatic annotation method. The method detects objects from driving images. Then, it extracts risk factors from the detected objects and automatically annotates captions as ground truth. The object category, position, and distance are important cues in extracting risk factors while driving. Hence, the method extracts risk factors by considering these cues. As an object detector, a faster R-CNN [10] trained with a COCO dataset [11] is used, and the detection threshold is set to 0.9.

TABLE I
RULE OF ANNOTATIONS.

priority	state
1	person/people crossing road
2	person/people on the sidewalk
3	detected traffic light
4	detected traffic sign
5	many cars parked on the shoulder of the road
6	object nearby

The following five classifications are selected as object categories from 80 in the COCO dataset: people, cars, bicycles, stop signs, and traffic rights. The objects are extracted corresponding to these classifications from detected objects using the faster R-CNN. Then, the bottom center of a bounding box is set as a reference point, as shown on the left side of Fig. 5. The Hough transform is applied to the image, and the point where detected lines intersect is set as a vanishing point. The reference and vanishing points give the direction attribute, i.e., left, right, or center, to each object, as shown in Fig. 5. Also, the distance from the vehicle to the objects is computed using the area of the bounding box. The thresholds are set for each category, and the distance attribute, i.e., normal, nearby, or far, is given. The given direction and distance attributes enable a risk factor attribute to be assigned to each of the objects. Note that the risk factor attributes refer to dangerous situations while driving, e.g., a pedestrian crossing a road in front of the vehicle. In this work, we set the six risk factor attributes shown in Table I. These attributes have priorities, and two attributes having higher priorities are given.

Next, correct captions are automatically annotated for each risk factor using the given attributes. Figure 6 shows examples of the annotated captions in our method. Three captions are annotated for the attribute of risk factor attribute with the highest priority, and two are annotated for that with the second highest priority. If the risk factor attribute of an image is only priority 6, we annotate three captions of it and two captions generated by the NBT model trained by the COCO dataset. During an evaluation, we use the three objects having the highest priority risk factor attributes.

To create the dataset, we collected 30,320 images and automatically annotated captions using the method, totaling 151,600. In our experiments, we used 25,987 images for training and 4,333 images for the evaluation.

B. Image captioning applying an attention mask

Our captioning method enables the model to generate captions focusing on only specific object regions. To this end, we control the weight of attention mechanism a_t for each object candidate region detected by RPN by mask processing. Figure 7 shows the overview of the proposed method.

An attention mask $\mathbf{A}_t$ is a one-hot vector having a number

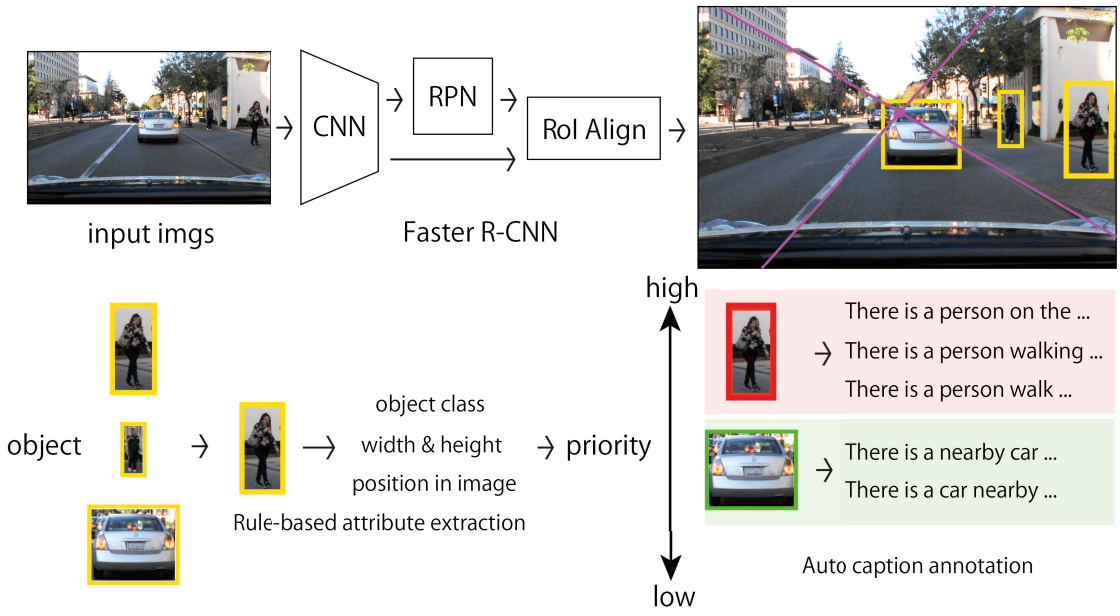

Fig. 4. Overview of our automatic annotation method.

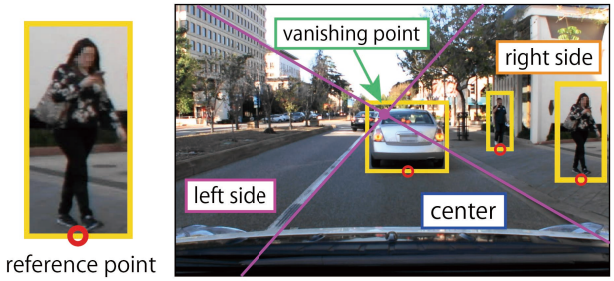

Fig. 5. Example of annotation rules.

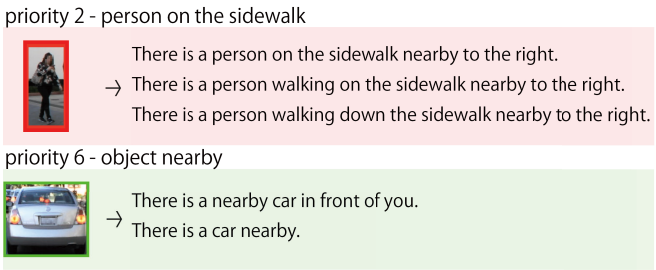

Fig. 6. Examples of automatically annotated captions.

of elements that correspond to the number of detected objects. The element corresponding to the object used to generate a caption is set to 1. On the basis of Eq. (6), c_t of the proposed method is calculated using the weighted sum of the attention mask and features of object candidate regions v_i:

$$c_t = \sum_{i=1}^{N} v_i \cdot a_{ti} \cdot \mathbf{A}_{ti}. \tag{17}$$

Moreover, unlike another attention-based models, NBT uses an attention model based on two LSTM layers. This can apply the attention mechanism for each object candidate region. Focusing on this network structure, our method can generate captions focusing on risk factors using v_t, on which we want to focus as an attention mechanism.

IV. EXPERIMENTS

Herein, we present an evaluation of the performance of our captioning method with our created dataset. To evaluate the captions generated, we used a questionnaire and automatic evaluation metrics, i.e., BLEU [12] and METEOR [13].

To demonstrate the effectiveness of our created dataset, we evaluated the captioning results using a questionnaire. We show two captions: one was generated by the NBT model trained using the COCO dataset, and the other was generated by the NBT model trained using our created dataset. Then, evaluators selected an answer from these three on which caption was suitable: “caption A,” “caption B,” or “ both.” Finally, we calculated the precision of each method, meaning the ratio selected by the evaluators. The questionnaire was conducted with 59 evaluators, and they were divided into five groups. We prepared 20 images and the corresponding captions for each group, totaling 100 images, and had them answer the aforementioned question for each.

To demonstrate the effectiveness of our captioning method, we evaluated the performance of the conventional NBT and our captioning method using automatic evaluation metrics: BLEU and METEOR. The models were trained using our dataset. Our method generated three captions by applying an attention mask based on the priorities of the risk factors decided by the rule-based annotations. Next, we computed the metrics with three captions of higher priorities in labeled captions as references.

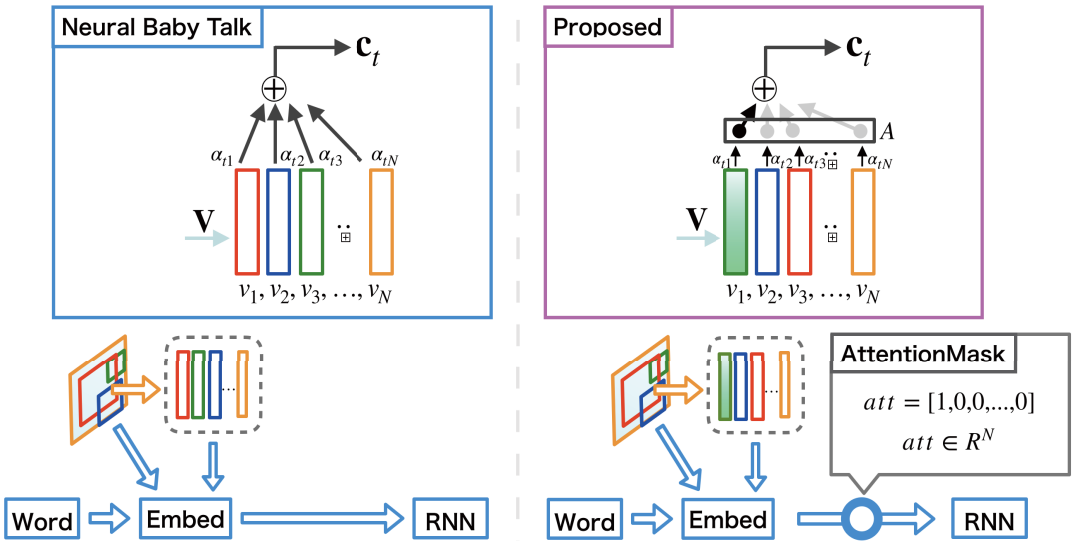

Fig. 7. Mask processing to an attention mechanism.

TABLE II
PRECISIONS WITH QUESTIONNAIRE

Method	Questionnaires' Precision
NBT	43.1%
Our NBT (priority 1 only)	**63.2%**

A. Training

We trained the NBT model using our dataset introduced in Sec. III-A. We used an RPN model trained using the COCO dataset as the object detection module in NBT. We set the number of units of attention and language LSTMs in the NBT language model to 512. First, we trained the NBT model with the COCO dataset, and then we fine-tuned it with our created dataset. We trained it with an Adam optimizer for 15 epochs, and we set the learning rates of the CNN and LSTM to 0.00001 and 0.0005, respectively. We set the mini-batch size to 10.

B. Evaluation results with questionnaire

Table II shows the results of the questionnaire. The precision of our method was 20 percent higher than that of the conventional method. Because the proposed method generates captions including category, position, and distance with respect to risk factors, its precision increased. Moreover, a comparison of each method's precision revealed that our created dataset is suitable for image captioning to get the driver's attention when needed.

C. Evaluation results with automatic evaluation metrics

Table III shows the evaluation results with BLEU and METEOR. Our method, which considers the priorities of risk factors, outperformed the conventional method in this evaluation. The results show that our method performed better than the conventional method in generating captions because it considered priorities and focused on each risk factor.

D. Comparison with generated captions

Figure 8 shows examples of captions generated by the conventional NBT model and our method. In the top-left of Fig. 8, the proposed method generates "There is a person on the sidewalk nearby to the right." for a woman crossing a road as the result of priority 1. The reason is that this result successfully includes words that indicate the appropriate classification, distance, and position. In contrast, the conventional NBT generates "A street with a lot of traffic on it.", which considers the entirety of the scene but is inadequate for attention attraction. These results demonstrate that our method can generate a caption suitable to get the driver's attention, enabling improved safety.

V. CONCLUSION

In this study, we created a dataset to get a driver's attention while operating a vehicle and generated multiple captions focusing on each risk factor. The proposed method can automatically create the dataset using object detection from images and rule-based annotation. Moreover, we achieved captioning for risk factors by introducing an attention mask to the conventional NBT model. Our future work includes designing a method for automatically constructing more suitable datasets and an appropriate captioning method for attention attraction so that drivers can deal with more dangerous situations.

REFERENCES

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[2] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

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

TABLE III

EVALUATION RESULTS WITH BLEU AND METEOR.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR
NBT (only priority1)	62.9	56.6	50.4	46.7	35.1
Ours (only priority 1)	**71.0**	**66.3**	**62.4**	**59.4**	**40.0**

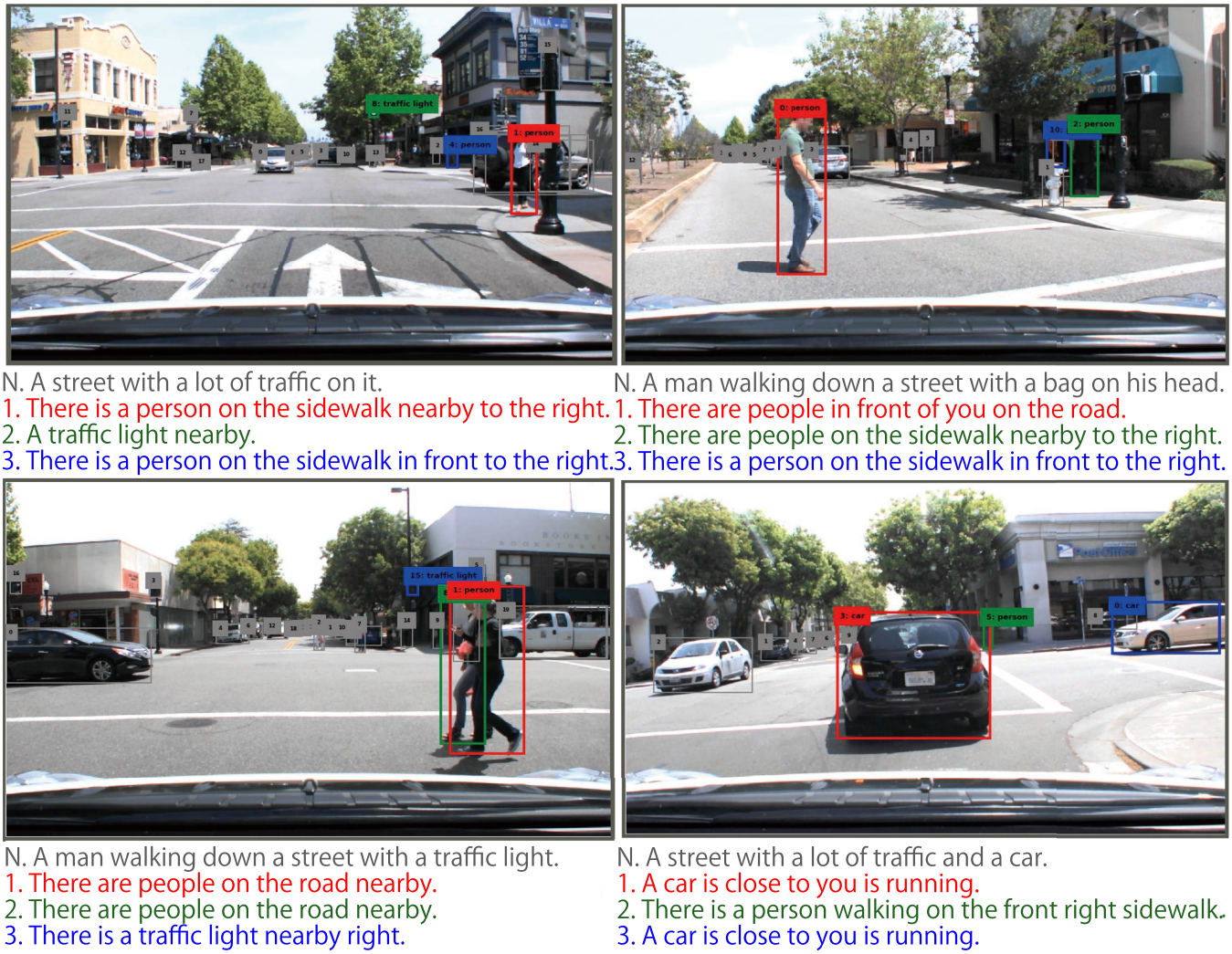

Fig. 8. Examples of generated captions. N is the captions generated by the NBT model. 1, 2, and 3 are captions generated by our method, where each number indicates the priority. The colorized bounding boxes in the images are risk factors corresponding to the same color captions.

[4] 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.

[5] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," in *Computer Vision and Pattern Recognition*, 2017, pp. 3242–3250.

[6] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Computer Vision and Pattern Recognition*, 2018, pp. 7219–7228.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer Networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2692–2700.

[9] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Annual Meeting of the Association for Computational Linguistics*, 2002.

[13] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in *EACL Workshop on Statistical Machine Translation*, 2014, pp. 376–380.