

The Traffic Scene Understanding and Prediction Based on Image Captioning

Publisher: IEEE

Cite This

PDF

Wei Li ; Zhaowei Qu ; Haiyu Song ; Pengjie Wang ; Bo Xue All Authors

17
Cites in
Papers3322
Full
Text Views
[Open Access](#) [Comment\(s\)](#)
Under a [Creative Commons License](#)

Abstract

Document Sections

I. Introduction

II. Related Works

III. Model

IV. Experiments

V. Conclusion

Authors

Figures

References

Citations

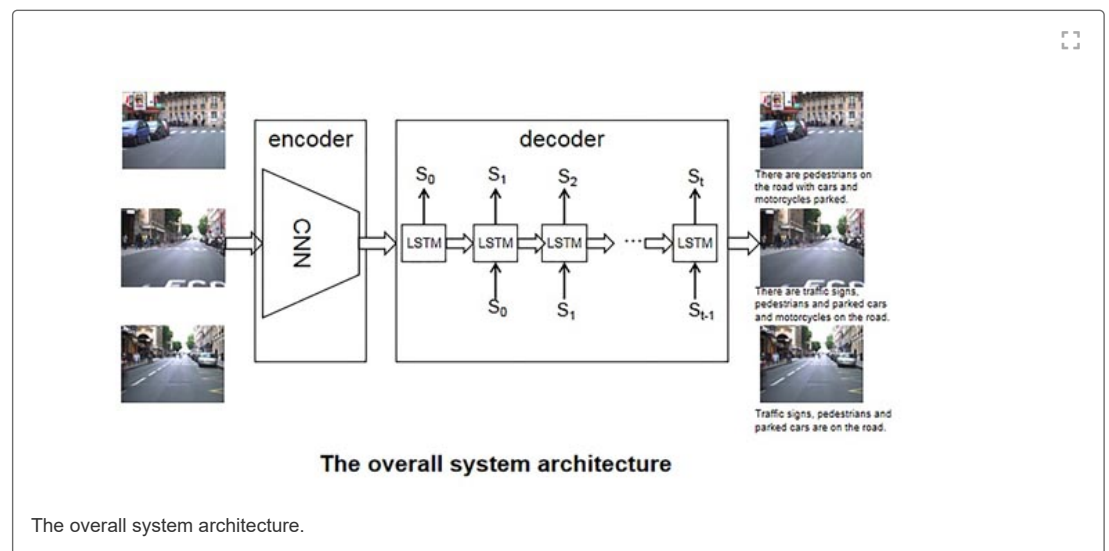
Keywords

Metrics

More Like This

Abstract:

The traffic scene understanding is the core technology in Intelligent Transportation Systems (ITS) and Advanced Driver Assistance System (ADAS), and it is becoming increasingly important for smart or autonomous vehicles. The recent methods for traffic scene understanding, such as Traffic Sign Recognition (TSR), Pedestrian Detection, and Vehicle Detection, have three major shortcomings. First, most models are customized for recognizing a specific category of traffic target instead of general traffic targets. Second, as for these recognition modules, the task of traffic scene understanding is to recognize objects rather than make driving suggestions or strategies. Third, numerous independent recognition modules disadvantage to fusing multi-modal information to make a comprehensive decision for driving operation in accordance with complicated traffic scenes. In this paper, we first introduce the image captioning model to alleviate the aforementioned shortcomings. Different from existing methods, our primary idea is to accurately identify all categories of traffic objects and understand traffic scenes by making full use of all information, and making the suggestions or strategy for driving operation in natural language by using Long Short Term Memory network (LSTM) rather than keywords. The proposed solution naturally solves the problems of feature fusion, general object recognition, and low-level semantic understanding. We tested the solution on our created traffic scene image dataset for evaluation of image captioning. Extensive experiments including quantitative and qualitative comparisons demonstrate that the proposed solution can identify more objects and produce higher-level semantic information than the state-of-the-arts.

Published in: [IEEE Access](#) (Volume: 9)

Page(s): 1420 - 1427

DOI: [10.1109/ACCESS.2020.3047091](#)

Date of Publication: 24 December 2020

Publisher: IEEE

Electronic ISSN: 2169-3536

Funding Agency:

PDF

Help

CCBY - IEEE is not the copyright holder of this material. Please follow the instructions via <https://creativecommons.org/licenses/by/4.0/> to obtain full-text articles and stipulations in the API documentation.



SECTION I.

Introduction

Developing new Intelligent Transportation Systems (ITS) which take into consideration the socio-economical, environmental, and safety factors of the modern society is one of the grand challenges of the next century. Autonomous vehicles are promising for improving the safety and efficiency of transportation and mobility systems by providing vehicle control during normal driving. However, a human driver will still be expected to have control responsibilities for an automated driving car in the case of an emergency. Therefore, the most possible solution to ITS may be mixed autonomous vehicles and human-driven vehicles rather than fully automated without human operation. Therefore, the Advanced Driver Assistance System (ADAS) plays an important role in the Intelligent Transportation System. Nevertheless, decades of researches for ADAS are concentrated on traffic target detection and classification, rather than complicated traffic scene perception or driving strategic decision.

In summary, the current ADASs have two shortcomings. First, The methods mainly focus on specific problems ranging from traffic sign recognition [1]– [3] and vector detection [4]– [6] to pedestrian detection [7]– [9]. Up to the present, most attention has been paid to the researches on single or multiple target detection. However, there are various targets likely not belonging to the above ones in reality, and the ADAS will omit the new targets to put the driver in danger [58], [59]. Second, the current ADASs could not supply the driver intelligent driving suggestion and strategy of traffic scene ahead yet [55]– [57]. In contrast, the intelligent suggestion and appropriate strategies are capable of curtailing the response time of the driver to avoid the accident. In the area of deep learning, image captioning has attracted much more attention in recent years, which can provide a solution to the aforementioned problems. Image captioning can generate descriptions of the traffic situations and assist with ADAS under special traffic conditions. Third, the existing methods could not fuse all categories of information to represent comprehensive scenes or abstract semantic concepts.

In this paper, we provide a novel solution for ADAS by utilizing image captioning. To the best of our knowledge, this is the first work that introduces image captioning into ITS(AVs). The main goal of this work is to allow the development of a new generation of ADAS solutions where control could actually be shared between the automated vehicle and the driver. Our contributions are as follows:

- We propose a solution for traffic scene understanding with the image captioning method. Benefiting from image captioning, our method can not only outperform the traditional target detection but also comprehend and discover all the targets threatening to driving.
- Human mistakes due to inattention have been identified as the major crash-contributing factor by large-scale naturalistic studies [10], [11]. Our solution could call the driver's attention when he is distracted. Furthermore, the solution can provide an operation decision or strategy for drivers in urgent and limited time, which can reduce traffic accidents significantly.
- As there is no appropriate dataset about image captioning in the field of traffic, we create a dataset based on the LaRA dataset [51] and perform our experiments on the dataset.

SECTION II.

Related Works

A. Traffic Sign Recognition

Traffic Sign Recognition (TSR) is the main issue for a driver assistance system as it has a dual role to control the road traffic as well as warning and guiding the driver. The first research on traffic sign recognition can be traced back to 1987 when Akatsuka and Imai [12] attempted to make an early TSR system. The TSR systems are usually comprised of two parts: traffic sign detection and traffic sign classification. The regions of interest are first extracted in detection, and then identified correctly or rejected in classification. The methods that were used for the extraction of regions of interest also include the color-based technique, as in [13]– [16]. In recent years, some researchers have used different architectures and forms of Neural Networks for classification [17]– [20].

B. Pedestrian Detection

Pedestrian detection in autonomous driving provides an important guarantee for safe and stable operation of the intelligent driving system in complex and uncertain environments. Traditional algorithms include HOG features [21], SIFT feature [22], Haar feature [23], LBP features [24], SVM, the Deformable Parts Model

(DPM) [25], and pedestrian detection based on action features. With the rapid development of deep learning, the recent researches on pedestrian detection have become more and more prominent. The standard Regional Convolutional Neural Network (R-CNN) detection framework [26] was enhanced with region proposals and CNN classification as in Faster R-CNN [27].

C. Vehicle Detection

Vehicle detection is a very important task for traffic information analysis that can be used in traffic control and management to ensure a safe transportation system. The most recent studies on vehicle detection can be categorized into two groups [28]. The first is based on vehicle appearance, and the second is based on vehicle motion. The appearance-based strategies depend mainly on visual features, including vehicle symmetry, texture, edges, and color [29]–[32]. The concept of motion-based techniques is to extract the moving vehicles based on their motion characteristics that separate them from the background, such as optical flow, frame differences, and background subtraction [33]–[35].

In recent years, there is an increasing concern about vehicle detection. Considerable researches have been devoted to improving the training efficiency by combining the CNN with the circular pairwise classification for vehicle recognition [36].

D. Image Captioning

Being capable of feature representation, deep convolutional neural networks have achieved dramatic progress in object detection. Many famous region proposal-based methods, such as R-CNN [26], Fast R-CNN [37], and Faster R-CNN [27] mentioned above, have great performance in object detection. In recent years, The regression-based single networks include YOLO [38], YOLO2 [39], SSD [40], and these types of detection methods use a single convolutional network to predict bounding boxes and class probabilities simultaneously from an input image. Although many successes have been achieved in traffic target detection, the categories of targets are specific and determined. However, there are many targets that the existing detectors can not recognize in the traffic environment. Furthermore, the previous methods of object detection can not supply the strategies and prediction for driving.

It is a relatively new task to let a computer use a human-like sentence to automatically describe an image. Image captioning will lead to a great number of possible applications, such as producing natural human-robot interactions, early childhood education, information retrieval, and visually impaired assistance, and so on. As a challenging and meaningful research field in artificial intelligence, image captioning is attracting more and more attention and becoming increasingly important. But this task is significantly harder than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community [41]. The image captioning models not only must be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, image captioning is viewed as a difficult problem and it is a very important challenge for machine learning and deep-learning algorithms. So there has been a recent surge of research interest in attacking the image captioning problems [42]–[45].

Recent works about image captioning mainly focus on language-based methods with an encoder-decoder framework [41], [46]–[50], where a convolutional neural network (CNN) encodes images into visual features, and a Recurrent Neural Network (RNN) decodes features into sentences [41].

Karpathy [48] proposed a system to provide natural language descriptions for image regions. He employed a CNN to extract features from image regions and an RNN to generate a short sentence for each region. Vinyals [41] also proposed a holistic method based neural image caption generator, known as NIC. He integrated a deep CNN as the image encoder for vision feature learning and an RNN for caption generating. Significantly higher BLEU scores were achieved in comparison with those state-of-the-art methods when evaluated using MSCOCO, Flickr8k, and Flickr30k datasets. Xu introduced an attention-based model automatically learning to describe the content of the image. He showed how the model was able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output caption [50].

SECTION III. Model

To apply the image captioning method in traffic scene understanding, we should focus on describing the factors influencing driving. Inspired by Vinyals [41], we introduce the image captioning method about the encoder-decoder model called Neural Image Caption (NIC). The encoder and the decoder are constructed by CNN and RNN respectively. The model is an end-to-end system that is fully trainable using stochastic gradient descent (SGD).

We start by briefly describing the encoder-decoder framework [41]. Given an image and the corresponding caption, the model directly maximizes the following objective.

$$\theta^* = \arg \max_{\theta} \sum_{(I,S)} \log p(S|I; \theta) \quad (1)$$

[View Source](#) 

where θ are the parameters of the model, I is an image, and $S = \{S_1, S_2, \dots, S_t\}$ is the correct caption. S represents any sentence and its length is unbounded. Using the chain rule, the log-likelihood of the joint probability distribution can be decomposed into ordered conditionals:

$$\log p(S|I) = \sum_{t=0}^N p(S_t|I, S_0, \dots, S_{t-1}) \quad (2)$$

[View Source](#) 

where N is the length of this particular example and we drop the dependency on model parameters for convenience. At training time, (S, I) is a training example pair, and the sum of the log probabilities is optimized as described in (2) over the whole training set using stochastic gradient descent (SGD).

To model $p(S_t|I, S_0, \dots, S_{t-1})$, it is natural to apply an RNN, where the variable number of words we condition upon up to $t-1$ is expressed by a fixed-length hidden state or memory h_t . This memory is updated after seeing a new input x_t by using a non-linear function f :

$$h_{t+1} = f(h_t, x_t) \quad (3)$$

[View Source](#) 

Though RNN has been proven successful on tasks such as speech recognition and text generation, it can be difficult to learn long-term dynamics being in part to the vanishing and exploding gradients problem. So in this stage, we utilize a Long-Short Term Memory (LSTM) net, which has shown state-of-the-art performance on sequence tasks such as translation. Formula (3) is replaced by LSTM as following:

$$h_{t+1} = LSTM(h_t, x_t) \quad (4)$$

[View Source](#) 

Our goal is to describe traffic scenes using image captioning based on the above framework, so we construct the architecture of our model (see Figure 1). Furthermore, as shown in Figure 1, abundant targets are described in the sentence to provide as much as possible information about the traffic environment.

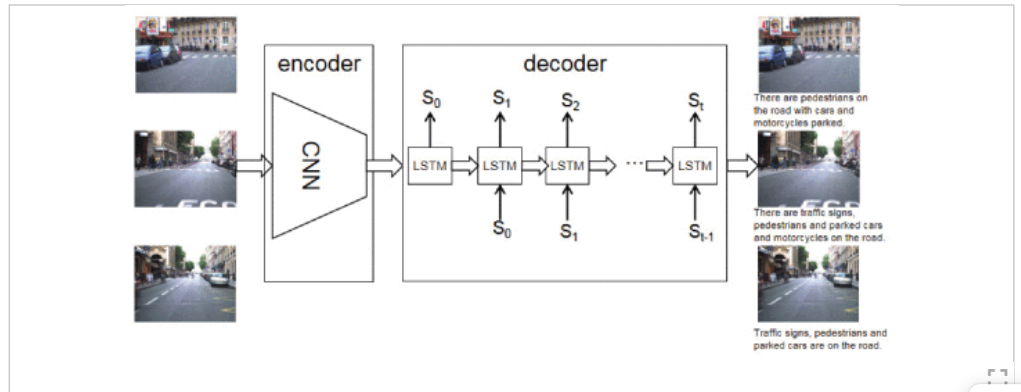


FIGURE 1.
The overall system architecture.

PDF

Help

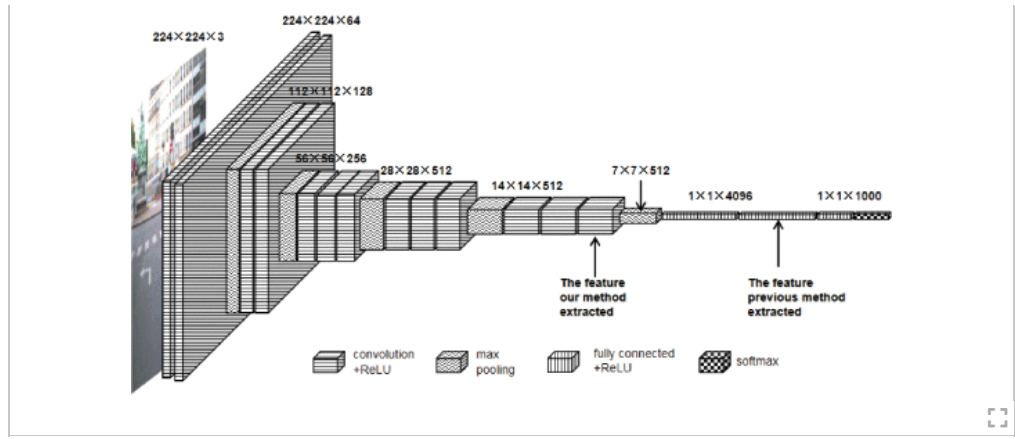


FIGURE 2.

The difference between our method and the previous methods in VGG16 network.

A. The Encoder

To encode images, we use a CNN architecture as a feature extractor. In the encoder stage of the image captioning model, we use the VGG-16 network [54] to extract the feature vectors of an image.

We extract features from a lower convolutional layer of the VGG-16 network to get the correspondence between the feature vectors and portions of the 2-D image. The work is different from the previous methods which instead used a fully connected layer [42], [44]. Thus, the decoder can selectively focus on a certain part of an image by selecting a subset of all the feature vectors.

B. The Decoder

The features extracted by the VGG16 network will be sent to the LSTM network to decode them and generate the corresponding image content description. The advantage of LSTM is modeling and predicting the implied context-dependency of information sequence. Figure 3 shows the basic memory unit structure of LSTM. The core of the LSTM model is a memory cell encoding knowledge at each time step for what inputs have been observed up to this step.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (5)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (7)$$

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (8)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \tanh(C_t) \quad (10)$$

[View Source](#)

where i_t , f_t , o_t , \tilde{C}_t , C_t , h_t are the input, forget, output, temporary memory, memory, and hidden state of the LSTM. The behavior of the cell is controlled by the gates which can either keep a value from the gated layer if the gate is 1 or zero value if the gate is 0. Furthermore, the forget gate is used to control whether to forget the current cell value, the input gate is used to read its input, and the output gate is used to decide whether to output the new cell value. Where σ is the sigmoid activation function. The LSTM can learn exceedingly complex and long-term temporal dynamics from these additional cells. Additional depth can be added to LSTMs by stacking them on top of each other, using the hidden state of the LSTM in layer $t-1$ as the input to the LSTM in layer t .

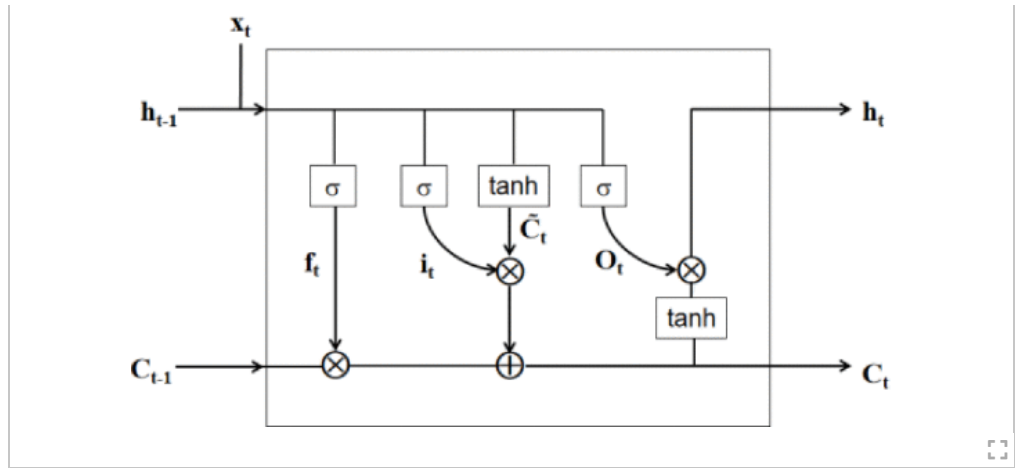


FIGURE 3.
The LSTM memory cell.

The unrolled LSTMs memories of the model are shown in Figure 4 and all LSTMs share the same parameters. The output h_{t-1} of the LSTM at time $t-1$ is fed to the LSTM at time t . All recurrent connections are transformed to feed-forward connections in the unrolled version as following:

$$x_{-1} = CNN(I) \quad (11)$$

$$x_t = W_e S_t, \quad t \in \{0 \dots N-1\} \quad (12)$$

$$p_{t+1} = LSTM(x_t), \quad t \in \{0 \dots, N-1\} \quad (13)$$

[View Source](#) ?

where W_e is word embedding and the image I is only input once at x_{-1} to inform the LSTM about the image contents. The loss function is the sum of the negative log-likelihood of the correct word at each step as follows:

$$L(I, S) = - \sum_{t=1}^N \log p_t(S_t) \quad (14)$$

[View Source](#) ?

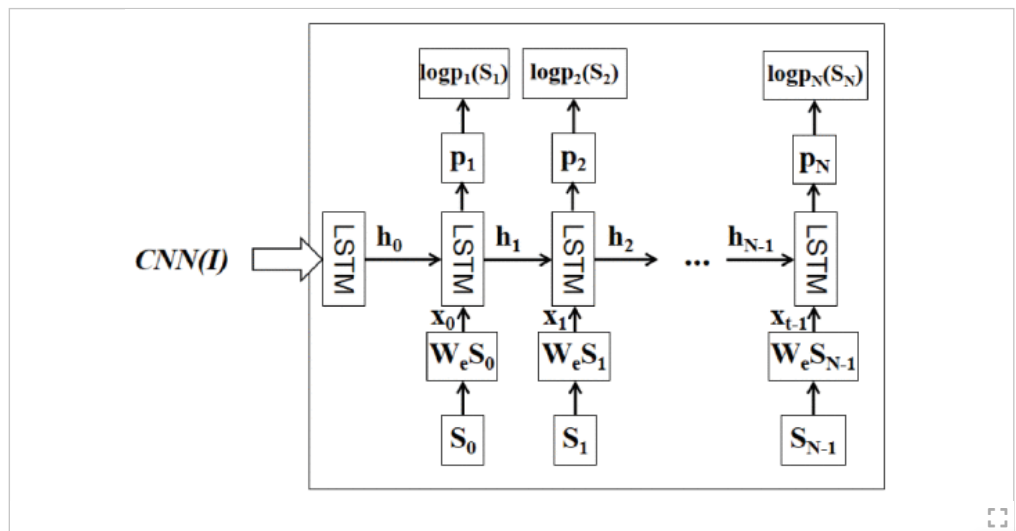


FIGURE 4.
The decoder framework.

The loss is minimized with respect to all the parameters of the LSTM, the top layer of the image embedder CNN and word embedding W_e (see Figure 4).









SECTION IV.

Experiments

A. Data Preparation

As the first work that introduces image captioning into traffic, to evaluate this work, we created an image captioning dataset for traffic scene understanding based on the LaRA dataset [51] and annotated each image with an English sentence. This dataset includes abundant traffic scenes with caption as a clue for the strategy or the prediction (see Table 1). We annotated 11178 images in the LaRA dataset manually and selected one-tenth of them randomly as the test set.

TABLE 1 The Samples of Our Created Dataset

Image	Caption	Image	Caption
	There is car ahead and some motors by the road.		There are some cars on both sides of the road, the traffic light is red, pedestrians on the road and a bus passing by.
	There are some motors and cars on both sides of the road.		There are traffic signs, pedestrians and parked cars and motorcycles on the road.
	There is a white car in the middle of the road and some cars on both sides of the road.		There is a car in ahead, some pedestrians are crossing the road and some cars in the distance, the light is red.
	There are sidewalks, cars, bicycles, billboards, traffic signs, red lights on both sides.		There are traffic signs, pedestrians, red traffic lights and cars on the road.

B. Training Detail

The experiments are implemented in Anaconda with PyTorch open-source machine learning framework. We choose the widely used VGG-16 network architecture as a feature extractor. The input images were set to 224×224 according to the VGG-16 network, the size of the feature map extracted from the VGG-16 network is 14×14 , and the dimension of the feature map is 512. The feature map was linearized into a vector which was the first input information into the following LSTM network.

In the training process, we applied the pre-trained VGG-16 network to extract the feature map of an image in the encoder. In the decoder, the LSTM knew the caption (ground truth) of the image, so that each LSTM cell had the source word and the destination word (the input and the output of an LSTM cell). The parameters of our image captioning model were trained with stochastic gradient descent using adaptive learning rate algorithms.

However, in the testing process, the difference with the training process was that the decoder didn't know the real caption of the image, and LSTM cells had no source word and destination word. The start of the caption was generated from the first LSTM cell based on the feature vector extracted from the VGG-16 network. The first word (S_1 in Figure 4) was generated base on the start of the caption and the second word (S_2 in Figure 4) was generated base on the first word. When the destination word is 'end' or the length of the generating caption exceeds the predefined threshold, the testing process can generate a caption of the input image.

In most cases, we must handle overfitting when training models. Nonetheless, we explored several techniques to deal with overfitting. The most obvious way to not overfit is to initialize the weights of the CNN component of our system to a pre-trained model (e.g., on ImageNet). We did this in our experiments and it did help quite a lot in terms of generalization. Another set of weights that could be sensibly initialized are W_e , the word embeddings. As no significant gains were observed after trying some initialization method such as mentioned in paper [52], we uninitialized the weights for simplicity.

We trained all sets of weights using stochastic gradient descent with a learning rate of 0.005. All weights were randomly initialized except for the CNN weight, which we trained it by ImageNet. After pretraining with ImageNet, we fine-tuned the model with our created dataset. We used 512 dimensions for the embedding and the size of the LSTM memory. We set the mini-batch size as 10.

C. Experimental Results

The most commonly used metric so far in the image description literature has been the BLEU (bilingual evaluation understudy) scores [53], which is a form of precision of word n-grams between generated and reference sentences. The BLEU evaluation metrics, considered as standard evaluation metrics, have been

widely used to evaluate image captioning methods in recent years. Following most previous works [41]– [43], we also used BLEU scores as evaluation metrics.

Since there is no open traffic dataset on image captioning, to verify the performance of our method, we performed the experiment on our created dataset. To further fairly evaluate our method, we compare ours with the baseline and state-of-art methods on the public dataset Flickr30K. As illustrated in Table 2, our method can achieve comparable performance compared with the two other methods.

TABLE 2 Performance of Our Method on Flickr30K Dataset in Terms of BLEU Scores

methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
NIC[41]	66.3	42.3	27.7	18.3
Att-NIC[43]	66.9	43.9	29.6	19.9
ours	64.7	46.4	31.2	22.4







We first trained the model on the Flickr30k dataset and fine-tuned the model on our dataset. The performance of our method could be enhanced if our dataset is optimized when it is annotated more suited for traffic scenes. We compared our method with the traffic target detection in ADAS on the same images. As shown in Table 3, in contrast to the existing works that only recognize a specific category of traffic targets, our proposed method not only recognizes general traffic targets affecting traffic factors but also generates a clue for driving suggestion or strategy. Apparently, the information generated by our method is much richer than that of traditional methods, such as TSR, Pedestrian Detection, Vehicle Detection.

TABLE 3 Qualitative Comparison With the Existing Methods of Traffic Scene Understanding

Raw Image	Pedestrian Detection	Vehicle Detection	Traffic Sign Recognition	Image Caption Detection
			Parking No Stopping Beware Pedestrian	There are Parking sign and pedestrians on the road with cars and motorcycles parked.
			Parking	There are traffic signs, pedestrians, parked cars, and motorcycles on the road.
			No Stopping Beware Pedestrian	There are pedestrians, traffic signs, and parked cars on the road.
			No Passing No Stopping	There are cars, traffic signs, sidewalks, and pedestrians on the road.

Furthermore, our model can identify more abundant semantic information from the traffic scene (see Table 4). For example, the textual caption generated by our method contains the relative position of traffic targets and the risk factors on the road. Thus, the caption can provide more information for the driver. The results show that the method is more valuable with the abundant semantic information of the traffic scene.

TABLE 4 The Examples of Image Captioning Generated by Our Method

Test image	Caption	Test image	Caption
	There are some cars on both sides of the road.		There are some cars, some motors and some pedestrians crossing the road.
	There are some motors under the parking sign.		There are two cars, a motor ahead, some pedestrians are crossing the road and the traffic light is red.
	There are some cars on the right sides of the road.		There are some motors, some pedestrians and a zebra crossings ahead.

From the results, obviously, our image captioning method can describe general traffic targets and more abundant semantic information at the same time. Since the image caption information is a clue for strategy or suggestion when operating a car, our method can be directly utilized in ADAS. Especially for some emergency conditions, the direct operation strategy or suggestion generated by image captioning is significant for the driver to avoid traffic accidents.

SECTION V.

Conclusion

The ITS is an active topic in the fields of traffic and artificial intelligence. However, most works focus on only recognizing certain targets involved in driving rather than providing driving suggestions. Image captioning originates from Machine Translating [60] and Natural Language Processing(NPL) [43], and mainly applied in human-robot interactions [61], early childhood education [62], information retrieval [63], and visually impaired assistance [64]. In this paper, we attempt to introduce the image captioning method, the new emerging technology in artificial intelligence, to ITS so as to generate high-level semantic information for driving suggestions or strategies. To train the model for ITS, we created a traffic image dataset for image captioning by associating sentences describing traffic scenes and providing driving suggestions with each image. The experiments on the traffic image dataset demonstrate its effectiveness, which we compare our method with traditional traffic target detection and recognition. The experiment results suggest that this method is a promising solution to ITS.

Currently, there is little research on describing general traffic targets and generating driving suggestions simultaneously. Furthermore, no previous study has been found introducing image captioning to ITS. As shown in our experiments, image captioning gives a new perspective on ITS and has great potential for ITS. However, directly applying the existing model of image captioning to ITS is not an optimal solution. It would seem, therefore, that further research focusing on specific model for the traffic domain is required in the future. In addition, the large-scale image dataset for image captioning needs collecting in ITS.

We will further improve the performance of image captioning in the following research work so that the strategy and prediction of driving could be more accurate. We will investigate how to collect and annotate a larger traffic dataset suitable for future traffic image captioning research.

Authors	▼
Figures	▼
References	▼
Citations	▼
Keywords	▼
Metrics	▼

ALSO ON IEEE XPLORE

Arabic Lip Reading With Limited Data Using ...

7 months ago · 1 comment

Two main challenges faced by deep learning systems are related to the amount ...

A Multifaceted Vision of the Human-AI ...

2 months ago · 1 comment

Human-AI collaboration has evolved into a complex, multidimensional ...

Identifying and Solving a Vulnerability in the ...

3 months ago · 1 comment

In Estonia, nationwide internet voting (i-voting) has been in use since 2005, ...

A Method for DDoS Attacks Prevention ...

8 months ago · 1 comment

Distributed Denial-of-Service (DDoS) attacks are among the most common ...

A TinyDL Model for Gesture-Based Air ...

10 months ago · 1 comment

The application of tiny machine learning (TinyM in human-computer ...

G

Start the discussion...

LOG IN WITH



OR SIGN UP WITH DISQUS



Name

Email

Password

By clicking submit, I authorize Disqus, Inc. and its affiliated companies to:

- Use, sell, and share my information to enable me to use its comment services and for marketing purposes, including cross-context behavioral advertising, as described in our [Terms of Service](#) and [Privacy Policy](#)
- Supplement the information that I provide with additional information lawfully obtained from other sources, like demographic data from public sources, interests inferred from web page views, or other data relevant to what might interest me, like past purchase or location data
- Contact me or enable others to contact me by email with offers for goods and services (from any category) at the email address provided
- Process any sensitive personal information that I submit in a comment for the purpose of displaying the comment
- Retain my information while I am engaging with marketing messages that I receive and for a reasonable amount of time thereafter. I understand I can opt out at any time through an email that I receive. Companies that we share data with are listed [here](#).



Share

Best Newest Oldest

Be the first to comment.

IEEE Personal Account

CHANGE
USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS
VIEW PURCHASED
DOCUMENTS

Profile Information

COMMUNICATIONS
PREFERENCES
PROFESSION AND
EDUCATION
TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800
678 4333
WORLDWIDE: +1 732
981 0060
CONTACT & SUPPORT

Follow

