Video Captioning Based on Both Egocentric and Exocentric Views of Robot Vision for Human-Robot Interaction

Soo-Han Kang¹ · Ji-Hyeong Han¹

Accepted: 15 October 2021 / Published online: 30 November 2021
© The Author(s) 2021

Abstract

Robot vision provides the most important information to robots so that they can read the context and interact with human partners successfully. Moreover, to allow humans recognize the robot's visual understanding during human-robot interaction (HRI), the best way is for the robot to provide an explanation of its understanding in natural language. In this paper, we propose a new approach by which to interpret robot vision from an egocentric standpoint and generate descriptions to explain egocentric videos particularly for HRI. Because robot vision equals to egocentric video on the robot's side, it contains as much egocentric view information as exocentric view information. Thus, we propose a new dataset, referred to as the global, action, and interaction (GAI) dataset, which consists of egocentric video clips and GAI descriptions in natural language to represent both egocentric and exocentric information. The encoder-decoder based deep learning model is trained based on the GAI dataset and its performance on description generation assessments is evaluated. We also conduct experiments in actual environments to verify whether the GAI dataset and the trained deep learning model can improve a robot vision system

Keywords Human robot interaction · Deep learning · Video captioning · Egocentric video

1 Introduction

As robots receive much more attention as partners for humans, robot technology, especially human-robot interaction (HRI), is rapidly developing [1–3]. HRI research covers various fields, including robot hardware control [4,5], robot navigation [6,7], and robot vision [8,9]. Human visual perception is our most important sense by which to understand situations and to interact with others. Likewise, vision information is very useful for robots to ensure natural and effective HRI. Therefore, robots must process and understand visual input properly. However, robot vision data does not simply consist of still images, but is in fact, video data. Thus, robots must recognize the context from video data while also con-

sidering the time flow. Video data analysis is an area that has long been studied as part of the field of computer vision. Currently deep learning-based approaches show the best performance.

Areas of interest in deep learning-based video analysis research consist of action recognition, video object tracking, video classification, and video description, among others. Video description is an especially complex problem because to goal is to generate the natural language sentences to describe a video. Despite its difficulties, video description has been studied actively because it is the best way to deliver detailed information about a video to humans [10,11]. The basic approach for deep learning-based video description is the encoder-decoder architecture, which encodes sequential video frames and decodes them to generate the description sentence. Typical models are the sequence-to-sequence (Seq2Seq) model [12], which uses a convolution neural network (CNN) and a recurrent neural network (RNN), and the transformer-based model [13], which shows the best performance in many fields especially neural machine translation.

The best way to ensure comprehension by the robot of its human partner is also to use natural language. Here, we define the problem as video description; however, there is a difference between robot vision data and general video data. With

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12369-021-00842-1>.

✉ Ji-Hyeong Han
jhhan@seoultech.ac.kr

Soo-Han Kang
shkang@seoultech.ac.kr

¹ Seoul National University of Science and Technology, Seoul, South Korea

Fig. 1 Example of the difference between the first person point of view and the third person point of view. The video is identical but it is interpreted in different ways based on the point of view

regard to robots, robot vision is basically egocentric vision; in other words, robot vision is a first-person perspective on the robot's side. The direction of interpretation for egocentric video data is slightly different from that in a general video analysis, as there is a subject called 'me' in egocentric video. For example, suppose there is a video taken by me as I walk down a street and another man is walking towards me from the opposite direction, as shown in Fig. 1. If the video is considered generally as the third person omniscient point of view, the video would then be interpreted as 'one person is walking down the street.' However, if the video is considered as the first-person egocentric point of view, the video would then be interpreted as 'I'm walking down the street' or 'A man is walking towards me.' To conclude, as robot vision is the first-person egocentric view from the robot's side, it is necessary to consider the egocentric view difference to generate robot vision descriptions.

To solve this problem, we propose a way to integrate the egocentric information and the exocentric information. Although interpretations of egocentric and exocentric video differ, the main model can be utilized identically. Accordingly, we apply encoder-decoder architecture deep learning models based on LSTM and transformer and utilize tokens to express the two different types of information at the same time, as illustrated in Fig. 2. To represent the first-person and the third-person points of view simultaneously, we construct a new dataset which is called GAI (Global, Action, and Interaction) dataset. The dataset consists of several egocentric view videos taken by a human with a wearable camera. We generate new description sentences for each video clip in three parts. The 'Global' part represents the exocentric information, the 'Action' part expresses the egocentric information, and the 'Interaction' part indicates both types (egocentric and exocentric) of information. The GAI descriptions of each video clip is distinguished by tokens,

(a) Encoder-decoder architecture of bi-directional LSTM with attention

(b) Encoder-decoder architecture of transformer

Fig. 2 The applied deep learning models, which are the encoder-decoder architectures of a bi-directional LSTM with attention and a transformer, to understand egocentric video and generate descriptions in three situations, global, action, and interaction, simultaneously. GBOS, GEOS, ABOS, AEOS, IBOS, and IEOS are the tokens to indicate the beginning of the sentence (BOS) and the end of the sentence (EOS) for each global (G), action (A), and interaction (I) description, respectively

i.e. <GBOS>, <GEOS>, <ABOS>, <AEOS>, <IBOS>, and <IEOS>, which are the beginning of sentence (BOS) and the end of sentence (EOS) for global (G), action (A),

and interaction (I) descriptions. The new dataset with GAI descriptions allows the user to apply the descriptions selectively based on the applications. Furthermore, it allows the video to be interpreted from multiple perspectives simultaneously. The core concepts of the GAI dataset are summarized below.

- The description sentences are in the order of global-action-interaction.
- The global sentence explains the overall situation, including details such as the place, light, and weather, among other information.
- The action sentence explains what the subject, i.e., the person, is doing.
- The interaction sentence explains the interaction situation or the behavior between the subject, i.e., the person, and others.
- Interaction can occur, or not, depending on the video clip.

Based on the above-mentioned method, it is possible to acquire as much information as possible by viewing the egocentric video differently depending on the situation. To the best of our knowledge, we are the first to propose a method that simultaneously considers egocentric and exocentric aspects when construing a first-person video. Our implementation and dataset are available on GitHub (<https://github.com/KangSooHan/GAI>).

We conduct experiments to demonstrate that deep learning models, which consist of an encoder for interpreting the video and a decoder for generating description sentences, train the egocentric and exocentric information simultaneously based on the proposed GAI dataset. To show the robustness of the proposed GAI dataset, we train different combinations of encoder and decoder deep learning models based on LSTM and transformer, which are the most famous deep learning approach for sequence data. Table 1 shows the applied encoder-decoder deep learning models and Fig. 2 shows the architecture of base models we apply. The experimental results are evaluated based on the BLEU score [14], and we compare the results when the global, action, and interaction parts are considered both simultaneously and separately. Also, we conduct experiments in an actual environment to demonstrate that the GAI dataset and the trained

models based on this dataset function properly for a robot vision system.

This paper is organized as follows. Section 2 provides a brief introduction to related research. In Sect. 3, we explain our main idea, i.e., GAI. Section 4 describes the experiments on the GAI dataset and discusses the experiment results. Finally, concluding remarks follow in Sect. 5.

2 Related Works

A description generation of egocentric video, which is the problem we want to solve in this paper, is an extension of video-captioning problems. Video-captioning problems are highly challenging, and many studies have been done to generate the most human-understandable descriptions, especially the deep learning-based approach. After the encoder-decoder model for sequential data was proposed, the deep learning-based approach has been applied successfully to the video-captioning problem. This approach is classified into three broad categories.

The first uses both a CNN and an RNN to interpret the images and sequential information of the video [12,15–17]. The encoding step in this approach applies pretrained CNN to obtain the features for each frame of the video, after which the generated features are passed to the RNN as inputs and the RNN creates outputs by considering the sequential information. After the encoding step, the decoding step generates the descriptions through another RNN, which takes the outputs of the encoder RNN as the inputs. This vanilla encoder-decoder model was improved by applying an attention mechanism. Bin et al. proposed the encoder-decoder model with the attention to represent which images affect which words and the extent to which they do so [15]. This approach has been widely used for video-captioning problems due to its rapid inference.

The second category uses only a CNN for the encoder [18,19]. One problem with an RNN is that it loses the information when the sequence is long. Thus, this method only uses a CNN to extract the features of the video frames. Additionally, sequential information is also represented by a CNN. This method shows better performance because it loses less information from the previous time steps and uses information regarding the relationship between neighboring frames.

The last category uses a transformer model, showing the best performance in neural machine translation [13]. Instead of using a CNN and an RNN, the transformer-based method applied a simple fully connected layer with self-attention for the encoder. This method outperforms previous methods in the literature using a CNN and an RNN.

The aforementioned research on video-captioning tasks has dealt with videos with general third-person omniscient points of view video. However, a few studies have concen-

Table 1 Applied encoder-decoder deep learning models

Model	Encoder	Decoder
L2L [12]	LSTM	LSTM
BiL2LAtt [15]	Bi-directional LSTM	LSTM with Attention
BiL2T	Bi-directional LSTM	Transformer
T2T	Transformer	Transformer

trated on the first-person egocentric point of view video, especially for video-captioning tasks [20,21]. Bolaños et al. created the egocentric video data with descriptions, but it was solved in the same manner as third-person point-of-view video data [21]. They created a dataset from video taken by a person with a wearable camera and wrote the descriptions only from that person's point of view. However, egocentric video contains much more information. Accordingly, we propose a method that uses all of the information from egocentric video to understand the video.

3 Methods

In this section, we introduce the method proposed here to build an egocentric video description dataset considering global, action, and interaction situations simultaneously. Many studies have attempted to understand videos with methods such as action recognition [22,23] and video description [24–26], among others. The video description task is the most complicated problem because it is necessary to understand the video context and generate the natural language sentences. Despite these difficulties, video descriptions are the most useful way to deliver the video information to humans due to the natural language advantage. In HRI, this is especially important because human partners need to understand the robot and natural language is the preferred way to deliver information to humans.

In HRI, robot vision is not like general videos taken by others because it uses the first-person egocentric point of view from the robot's side. Previous research on robot vision typically focuses on understanding and interpreting others' behaviors in robot vision [27,28]. However, as noted earlier, egocentric video consists of various types of information, not merely others' information. As humans, we can see what we are doing and also what is happening in front of us through our own eyes. In fact, there is a strong relationship between what one is doing and what is happening. For example, if a person sees a soccer field and teammates, most of that person's behaviors will involve running and/or kicking a ball. Thus, if we only focus on what one is doing or what is happening in egocentric view video, we will lose information from a different point of view, and this means we disregard the correlation between the different types of information.

To solve this problem, we describe egocentric video in three situations: global, action, and interaction. The global description explains the circumstances shown in the video to represent exocentric information. For both the human and the robot, because the global situation is the basis on which to determine their behaviors and to understand the context, the global description is the most important aspect, and it consists of as many details as possible to represent the context of the video. The action description represents what the

person is doing, similar to a general egocentric video analysis. The action description focuses on the person's behavior, i.e., an egocentric view of the subject's behavior, rather than on detailed information of the context. In HRI, the action description is very useful because it describes what the robot is doing and conveys the robot behavior information to the human partner. The interaction description represents the interaction between the subject of the egocentric view, i.e., the subject of the action description, and others in the video, i.e., others in the global description. Thus, the interaction description is a combination of egocentric and exocentric views. In HRI, the interaction description is the basis by which to understand the interaction between humans and robots and to determine whether or not the current interaction context is proper. Unlike the global and action descriptions, the interaction description may or may not be in the video because it depends on whether or not the subject interacts with others in the video. Using the proposed GAI dataset, the egocentric videos are interpreted in a simultaneous manner rather than only considering global, action, or interaction situations independently.

To build an organized GAI dataset, we define certain rules to follow. The first is the order of the descriptions. The order of descriptions in the GAI dataset is global, action, and then interaction descriptions. There are two reasons for our descriptions structure. First, the global description contains the most detailed information of the video. Second, many deep learning models that predict subsequent words are based on generated sentences (see Appendix A). In general, first-person view videos have the most information about the overall situation; therefore, the global description is placed first and the action descriptions follow. The interaction description is last because it should integrate the information from the global and action descriptions. Also, each description uses the same tense. Organizing descriptions in this way prevents unnecessary multiple appearances of information in each description and makes the relationship between situations more feasible.

The second rule is for the global descriptions. Global descriptions explain the overall context first and then add more detailed information, such as the place, weather, and surrounding environment information. Global descriptions cannot contain first-person pronouns such as I, my, me, or we, as they interpret the video from a third-person point of view.

The third rule is for the action descriptions. Because action descriptions explain what a person is doing in an egocentric video, they must contain first-person pronouns such as I, my, or me. Moreover, by considering that the video is egocentric, despite the fact that there is no explicit behavior by the subject in the video, the action descriptions must be built, such as "I'm standing still" and "I'm listening."

Table 2 Summary of egocentric video datasets [37]

Dataset	Activities	Content Details	GT Label and Task
UT ego [29,30]	Life-logging activities	11 events	Important region prediction and video summarization
GTEA [31,32]	Food preparation	7 types of food	Human action recognition and segmentation
GTEA Gaze [33]	Food preparation	30 kinds of food	Human action recognition and segmentation
CMU-MMAC [34]	Cooking	5 recipes	Human action recognition
JPL [35]	Social interaction	7 types of activities	Human action recognition
EGO-GROUP [36]	Social Interaction	4 different scenarios	Human pose estimation

The last rule pertains to the interaction descriptions. If there is no interaction between the subject of the video and others, the interaction description is then represented using a <None> token. If there is interaction between the subject of the video and others, the interaction description must contain words such as we or me and ‘my friend’ to represent the interactive subjects.

The GAI dataset provides several advantages, especially in the HRI field. The first is that it makes HRI more efficient because it explains egocentric robot vision in as great of detail as possible. Second, users can choose the situation from the global, action, and interaction types selectively based on their applications and can easily apply the dataset to existing models. Finally, it considers the global, action, and interaction situations in egocentric robot vision simultaneously.

4 Experiments

This section explains the GAI dataset and experiments based on the dataset.

4.1 GAI Dataset

Robots are usually designed similarly to humans. In particular, robot vision is very similar to human vision because it relies on egocentric vision from the robot’s side, identical to human vision. Thus, when robots learn how to understand visual information, egocentric videos should be used. There are two main types of egocentric video datasets. The first is focused on special applications such as cooking [38] or human behaviors [32], and the second consists of everyday life videos without a special objective. Table 2 summarizes the characteristics of the egocentric video dataset in the context of this study. Everyday life videos are more appropriate for robot learning to understand everyday situations. In particular, we want to use untrimmed video in order for robots

to achieve acceptable general performance. Therefore, we chose the UT egocentric dataset which contains four everyday life videos taken by a person using a Looxcie wearable camera without any constraints or special objectives. Each video in UT egocentric dataset is around three to five hours [29,30].

We built the GAI dataset based on the UT egocentric dataset. First, we divided the videos of the UT egocentric dataset into small clips which contain only one context. There are approximately 1,600 video clips in the GAI dataset, and each video clip has a different length. Then, we created two sets of GAI description sentences for each clip. The GAI description sentences were built by two labelers and the video clips were randomly allocated to them. The labelers were instructed in detailed and clear rules to build GAI description sentences as explained in Sect. 3. The built description sentences were cross-checked and the third person finalized the GAI descriptions.

The biggest difference and advantage between the GAI dataset and existing egocentric video datasets are the ground truth (GT) labels and target tasks. The typical tasks of egocentric video dataset are video summarization [29,30] and human action recognition and segmentation [31–36] as shown in Table 2. Thus, the existing egocentric video datasets have focused on recognizing what the human subject is doing by considering the egocentric aspect only and provided the simple ground truth labels for human action recognition. Unlike the existing datasets, the GAI dataset provides the natural language descriptions for three different perspectives, which are global, action, and interaction, to understand the egocentric videos in the context manner. For the best of our knowledge, the GAI dataset is the first dataset that explains first-person videos using natural language descriptions by considering egocentric and exocentric aspects simultaneously.

Figures 3 and 4 show the basic statistics of GAI dataset. Figure 3 shows statistics pertaining to the number of frames for each video clip. Figure 4 shows the word frequency statis-

Fig. 3 Statistics of the number of frames for the GAI dataset

Fig. 4 Statistics of word frequency for the GAI dataset

tics and 531 words in total are used in the GAI dataset descriptions. The most frequent words not counting articles and pronouns are related to the frequent situation in the videos, such as restaurants, marts, and situations in which people are eating. When the GAI dataset would be used for other applications and model training, these basic statistics would be useful information for efficient data preprocessing.

4.2 Model

We applied encoder-decoder based deep learning models to learn the GAI dataset. These models consist of an encoder which interprets the visual information and a decoder which generates the natural language descriptions. We used the pre-trained VGG16 as a backbone network to extract the image features from video frames, and the extracted features were used as inputs for the encoder to train the sequential information. The decoder generates the words sequence based on encoder output. To show the robustness of GAI dataset, we trained four different combinations of LSTM and trans-

former for deep learning models, i.e. L2L (vanilla LSTMs for encoder and decoder) [12], BiL2LAtt (bi-directional LSTM for encoder and LSTM with attention for decoder) [15], L2T (bi-directional LSTM for encoder and transformer for decoder), and T2T (transformers for encoder and decoder) models. Figure 2 shows the architecture of base models we applied, which are BiL2Att and T2T models, and Table 1 shows the summary of each model's encoder and decoder combination.

4.3 Experimental Results

In this subsection, we explain the experimental results of the GAI dataset. The experimental results were evaluated using the bilingual evaluation understudy (BLEU) score, which represents how well the description sentence is generated. For the interaction situation, there are two cases, with interaction and without interaction, and in this case the evaluation used the area under the curve (AUC). Table

Table 3 Experimental results when training GAI situations simultaneously. The global, action, and interaction descriptions are evaluated using a BLEU score, each averaged from BLEU@1 to BLEU@4, and the interaction existence classification is evaluated by AUC

	G(BLEU)	A(BLEU)	I(BLEU)	I(AUC)
L2L [12]	40.15	32.49	16.98	0.5243
BiL2LAtt [15]	43.17	35.71	21.53	0.5527
BiL2T	46.24	41.69	12.19	0.6011
T2T	46.09	42.73	19.08	0.6130

Table 4 Performance comparison of BiL2T and T2T models when considering GAI simultaneously and separately. The global, action, and interaction descriptions are evaluated in terms of BLEU score, which is averaged from BLEU@1 to BLEU@4, and the interaction existence classification is evaluated by AUC

Model	Types	G(BLEU)	A(BLEU)	I(BLEU)	I(AUC)
BiL2T	GAI	46.24	41.69	12.19	0.6011
	GA	43.59	42.97	–	–
	GI	46.88	–	28.61	0.6133
	AI	–	44.86	14.10	0.6785
	G	47.50	–	–	–
	A	–	44.36	–	–
	I	–	–	28.80	0.5416
T2T	GAI	46.09	42.73	19.08	0.6130
	GA	47.87	37.11	–	–
	GI	46.82	–	21.43	0.6067
	AI	–	41.57	23.89	0.6428
	G	47.06	–	–	–
	A	–	39.69	–	–
	I	–	–	16.17	0.5924

Fig. 5 Correct, partially correct, and incorrect examples of generated global, action, and interaction descriptions for the GAI dataset in the T2T model. The figures are key frames of video clips and there are

ground-truth descriptions and model-generated descriptions. The red descriptions represent the incorrect cases

3 shows the results of each trained model when considering global, action, and interaction descriptions simultaneously. As shown in Table 3, the results confirm that the L2L model without the attention method had the lowest performance, though for BiL2LAtt model with the attention, it performed better. Specifically, BiL2T and T2T, which used a transformer layer in the decoder, demonstrated significant performance improvements compared to other models. In addition, for BiL2T and T2T, the global BLEU scores showed similar performance capabilities at 46.24 and 46.09, respectively, with T2T showing slightly better performance in action and interaction situations.

Table 4 shows the results when considering GAI situations simultaneously and separately. We conducted experiments on the GAI situation in seven ways: considering three situations simultaneously (GAI), considering two situations simultaneously (GA, GI, AI), and finally considering each situation separately (G, A, I). These results show that there was no significant change in the performance when learning situations separately or simultaneously. However, with regard to interaction description, when training action and interaction descriptions at the same time (AI type) showed the best AUC score, and when training interaction description alone (I type) showed the worst AUC score. Furthermore, when considering GAI simultaneously, we found that the T2T performance

was slightly better than the BiL2T performance, but action and interaction descriptions of the L2T model were generally better than those of the T2T model. The reason for this result is that directional flow of video is important to capture the movements for action and interaction descriptions, thus BiL2T model which encoded the videos in directional manner using the bi-directional LSTM showed better performance in action and interaction descriptions than T2T model which encoded the videos without directional manner using the transformer.

Figure 5 shows the representative correct, partially correct, and incorrect description generation results of the trained T2T model based on GAI dataset. Moreover, to verify that the proposed GAI dataset and the trained model are feasible for robot vision, we conducted experiments in an actual environment. We used an Intel RealSense camera D435i and NVidia Jetson Xavier to implement an environment similar to that of a robot vision system. There were six situations which were set as similar situations in UT egocentric videos. Figure 6 shows the actual environment experimental results of BiL2LAtt and T2T models. As shown in Fig. 6, global and action descriptions were generated relatively well; however, the interaction descriptions were generated incorrectly. We discuss the causes of this result in the following subsection.

Fig. 6 The results of generated global, action, and interaction descriptions for the experiments in actual environments. The figures are key frames of videos taken from these experiments, and there are generated descriptions of the two different trained models, i.e., BiL2Latt

and T2T, when we consider the GAI situations simultaneously in both cases. The orange descriptions are the partially correct cases and the red descriptions are the incorrect cases

4.4 Discussion

The experiments showed similar performance outcomes when training the GAI situation simultaneously in the uni-model and when training the GAI situation separately with multiple models. This means that it is possible to generate a situation-aware model with fewer parameters. Even in the case of several interaction situations, training GAI simultaneously has been shown to lead to better performance compared to training the GAI separately. This result proves the benefit of the GAI dataset and our hypothesis that considering the GAI situation simultaneously would help to understand egocentric videos.

In the interpretation of the experimental results in interaction situations, we used the AUC score for comparison rather than using the BLEU score. Due to the interaction description statement, only one token <NONE> can appear. In this case, BLEU@1, which compares only one word, shows a high value. However, BLEU@234, which compare more than one word, shows very low scores, making BLEU score of the interaction descriptions difficult to trust. Therefore, we used AUC scores instead of BLEU scores to evaluate the interaction situation.

Additionally, our data are associated with a data balance problem. Because the UT egocentric videos were taken by a

person along with a friend, most of the interaction situations in the dataset contained the same people. Thus, the model learned not to classify the existence of interactions but to distinguish whether or not that friend existed. There are also very few descriptions of each interaction situation. To solve this problem, it is necessary to build a GAI dataset with more diverse interaction situations.

One possible robot application of the proposed GAI dataset is human-robot cooperation. When a robot engages in some tasks with a human partner, the robot must understand not only the current context in general but also what the robot is doing and what the two partners are doing at this moment. For example, we can consider a situation in which a robot and a human partner build blocks together among other many different toys. The robot must understand that the current context is ‘building blocks’ among other toys based on the global description. Then, the robot must figure out what it is doing at this moment based on the action description; for example, it selects a large block or a small block or waits until a human partner selects a block. Finally, the robot must understand what the human partner does at this moment based on the interaction description, such as when the human partner places a large block and waits for the robot’s turn, with the robot then deciding on its next move, such as selecting a small

block and placing it on a large block based on the interaction description.

5 Conclusion

This paper proposed a new approach by which to understand robot vision from an egocentric point of view for HRI, as robot vision can be considered to utilize the egocentric video from the robot's side. We constructed a new dataset referred to as the GAI dataset, consisting of global, action, and interaction descriptions for egocentric videos so as to understand the egocentric videos based on both egocentric and exocentric information simultaneously. We conducted experiments to demonstrate the robustness of the GAI dataset. Four different combinations of LSTM and transformer layers for the encoder and the decoder of video-captioning deep learning model were trained based on the GAI dataset. The experimental results verified that the descriptions were generated better when we considered GAI situations simultaneously as compared to when we considered them separately. Moreover, we conducted experiments in an actual environment to show that the GAI dataset and the trained model based on it were feasible for use on a robot vision system.

Funding This work was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. 2018R1C1B6007230).

Declarations

Conflicts of interest The authors declare that they have no conflicts of interest to report regarding the present study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

APPENDIX

Decoding Phase in a Video-Captioning Model

In this chapter, we explain the decoder phase in the encoder-decoder architecture of video-captioning model. In the

encoding phase, there are video frames $V = \{v_1, v_2, \dots, v_{n_v}\}$ with the maximum video frame number n_v which the pre-trained VGG-16 model is applied for each frame to extract features of each frame, $F = \{f_1, f_2, \dots, f_{n_v}\}$. Encoder information $E = \{e_1, e_2, \dots, e_{n_v}\}$ is generated through features F with encoder layer L_E (e.g., vanilla LSTM or a transformer). Finally, this generated encoding information E is passed to the decoder.

During the training phase, the decoder is trained through the teacher forcing method. The decoder uses E and $Y = \{y_1, y_2, \dots, y_{n_c}\}$ information and predicts $\hat{Y} = \{y_2, y_3, \dots, y_{n_c}, y_{n_c+1}\}$ through the decoder layer L_D . Y is the description provided by the video-captioning dataset and n_c is the maximum captioning time step. In this paper, L_D uses two layers; the first being the LSTM and the second being the transformer.

A. LSTM

The LSTM formula used in the model is as follows:

$$\begin{aligned} f_t &= \sigma(W_x^f x_t + W_h^f h_{t-1} + b^f) \\ i_t &= \sigma(W_x^i x_t + W_h^i h_{t-1} + b^i) \\ o_t &= \sigma(W_x^o x_t + W_h^o h_{t-1} + b^o) \\ j_t &= \lambda_j (W_x^j x_t + W_h^j h_{t-1} + b^j) \\ c_t &= c_{t-1} \odot f_t + i_t \odot j_t \\ h_t &= \lambda_h (c_t) \odot o_t \end{aligned}$$

The LSTM used in the decoder interprets video and word information. The vector for the input uses a concatenated word and video vector. One problem is that the time lengths of a word and a video differ. To solve this problem, the input of video vector $E' = \{e_{n_v+1}, e_{n_v+2}, \dots\}$ is created with zero-padding as input to LSTM instead of $E = \{e_1, e_2, \dots\}$. Thus, y_1 and E' are required to predict the first generative word y_2 and, repeatedly, y_{t-1} and E' can predict y_t at time t . This equation indicates that the first generated words are used to predict later words. Therefore, GAI methods, which predict global situations containing the most information first and then action and interaction situations, integrate the global, action, and interaction context efficiently.

B. Transformer

The Transformer formula used in the model is as follows:

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \dots, head_h)W, \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V), \\ Attention(Q, K, V) &= softmax(QK^T / \sqrt{d_k})V \end{aligned}$$

The information used in the transformer decoder is word and video information, as in LSTM. However, unlike LSTM, the transformer module does not use E' as direct input, instead using it in the middle of the module. In the attention formula, Q uses word vector Y and K and V use video vector E' . The transformer module also requires y_1 and E' to predict y_2 for the first word and, repeatedly, y_{t-1} and E' can predict y_t at time t . The transformer decoder produces richer and better words because it not only considers words in one direction but also predicts the next words with attention. Accordingly, GAI methods are useful in the transformer decoder as well to integrate the global, action, and interaction context.

References

- Kong Yu, Fu Yun (2018) Human action recognition and prediction: A survey. arXiv preprint [arXiv:1806.11230](https://arxiv.org/abs/1806.11230)
- McColl D, Hong A, Hatakeyama N, Nejat G, Benhabib B (2016) A survey of autonomous human affect detection methods for social robots engaged in natural hri. *J Intell Robot Syst* 82(1):101–133
- Ji Yanli, Yang Yang, Shen Fumin (2019) Heng Tao Shen, and Xuelong Li. A survey of human action analysis in hri applications, *IEEE Transactions on Circuits and Systems for Video Technology*
- Lunghi Giacomo, Marin Raul, Di Castro Mario, Masi Alessandro, Sanz Pedro J (2019) Multimodal human-robot interface for accessible remote robotic interventions in hazardous environments. *IEEE Access*, 7:127290–127319
- Ruiz Ariel Y Ramos, Rivera Luis J Figueroa, Chandrasekaran Balasubramaniyan (2019) A sensor fusion based robotic system architecture using human interaction for motion control. In: 2019 IEEE 9th annual computing and communication workshop and conference (CCWC), pages 0095–0100. IEEE
- Vasquez Dizan, Stein Procópio, Rios-Martinez Jorge, Escobedo Arturo, Spalanzani Anne, Laugier Christian (2013) Human aware navigation for assistive robotics. In: experimental robotics, pages 449–462. Springer
- Marques Francisco, Gonçalves Duarte, Barata José, Santana Pedro (2017) Human-aware navigation for autonomous mobile robots for intra-factory logistics. In: international workshop on symbiotic interaction, pages 79–85. Springer
- Moghadas M, Moradi, H (2018) Analyzing human-robot interaction using machine vision for autism screening. In: 2018 6th RSI international conference on robotics and mechatronics (IcRoM), pages 572–576. IEEE
- Liu Miao, Tang Siyu, Li Yin, Rehag James M (2020) Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: European conference on computer vision, pages 704–721. Springer
- Nguyen Anh, Kanoulas Dimitrios, Muratore Luca, Caldwell Darwin G, Tsagarakis Nikos G (2018) Translating videos to commands for robotic manipulation with deep recurrent neural networks. In: 2018 IEEE international conference on robotics and automation (ICRA), pages 1–9. IEEE
- Silvia Cascianelli, Gabriele Costante, Ciarfuglia Thomas A, Paolo Valigi, Fravolini Mario L (2018) Full-gru natural language video description for service robotics applications. *IEEE Robot Autom Lett* 3(2):841–848
- Venugopalan Subhashini, Rohrbach Marcus, Donahue Jeffrey, Mooney Raymond, Darrell Trevor, Saenko Kate (2015) Sequence to sequence-video to text. In: proceedings of the IEEE international conference on computer vision, pages 4534–4542
- Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, Polosukhin Illia (2017) Attention is all you need. In: advances in neural information processing systems, pages 5998–6008
- Papineni Kishore, Roukos Salim, Ward Todd, Zhu Wei-Jing (2002) Bleu: a method for automatic evaluation of machine translation. In: proceedings of the 40th annual meeting of the association for computational linguistics, pages 311–318
- Yi B, Yang Y, Fumin S, Ning X, Tao SH, Xuelong L (2018) Describing video with attention-based bidirectional lstm. *IEEE Trans Cybernet* 49(7):2631–2641
- Li Xuelong, Zhao Bin, Lu Xiaoqiang, et al (2017) Mam-rnn: Multi-level attention model based rnn for video captioning. In: IJCAI, p 2208–2214
- Bin Yi, Yang Yang, Shen Fumin, Xu Xing, Shen Heng Tao (2016) Bidirectional long-short term memory for video description. In: proceedings of the 24th ACM international conference on Multimedia, p 436–440
- Fang K, Zhou L, Jin C, Zhang Y, Weng K, Zhang T, Fan W (2019) Fully convolutional video captioning with coarse-to-fine and inherited attention. In: proceedings of the AAAI conference on artificial intelligence 33:8271–8278
- Liu Sheng, Ren Zhou, Yuan Junsong. (2020) Sibnet: Sibling convolutional encoder for video captioning. *IEEE Trans Pattern Anal Mach Intell*
- Fan Chenyou, Crandall David J (2016) Deepdiary: Automatically captioning lifelogging image streams. In: European conference on computer vision, pp 459–473. Springer
- Bolaños M, Peris A, Casacuberta F, Soler S, Radeva P (2018) Egocentric video description based on temporally-linked sequences. *J Vis Commun Image Represent* 50:205–216
- Tran Du, Wang Heng, Torresani Lorenzo, Ray Jamie, LeCun Yann, Paluri Manohar (2018) A closer look at spatiotemporal convolutions for action recognition. In: proceedings of the IEEE conference on computer vision and pattern recognition, p 6450–6459
- Lin Ji, Gan Chuang, Han Song (2019) Tsm: Temporal shift module for efficient video understanding. In: proceedings of the IEEE international conference on computer vision, p 7083–7093
- Wang Bairui, Ma Lin, Zhang Wei, Liu Wei (2018) Reconstruction network for video captioning. In: proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), June
- Lianli G, Zhao G, Zhang Hanwang X, Shen HT (2017) Video captioning with attention-based lstm and semantic consistency. *IEEE Trans Multimedia* 19(9):2045–2055
- Yang Y, Jie Z, Jiangbo A, Yi B, Alan H, Tao SH, Yanli J (2018) Video captioning by adversarial lstm. *IEEE Trans Image Process* 27(11):5600–5611
- Ryoo MS, Fuchs Thomas J, Xia Lu, Aggarwal Jake K, Matthies Larry (2015) Robot-centric activity prediction from first-person videos: What will they do to me? In: 2015 10th ACM/IEEE international conference on human-robot interaction (HRI), p 295–302. IEEE
- Koppula Hema S, Ashutosh S (2015) Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans Pattern Anal Mach Intell* 38(1):14–29
- Lee Yong Jae, Ghosh Joydeep, Grauman Kristen (2012) Discovering important people and objects for egocentric video summarization. In: 2012 IEEE conference on computer vision and pattern recognition. p 1346–1353. IEEE
- Lu Zheng, Grauman Kristen (2013) Story-driven summarization for egocentric video. In: proceedings of the IEEE conference on computer vision and pattern recognition, p 2714–2721

31. Fathi Alireza, Ren Xiaofeng, Rehag James M (2011) Learning to recognize objects in egocentric activities. In: CVPR 2011, p 3281–3288. IEEE
32. Li Yin, Ye Zhefan, Rehag James M (2015) Delving into egocentric actions. In: proceedings of the IEEE conference on computer vision and pattern recognition, pages 287–295
33. Fathi Alireza, Li Yin, Rehag James M (2012) Learning to recognize daily actions using gaze. In: European conference on computer vision, p 314–327. Springer
34. Torre Fernando De la, Hodgins Jessica, Bargteil Adam, Martin Xavier, Macey Justin, Collado Alex, Beltran Pep (2008) Guide to the carnegie mellon university multimodal activity (cmu-mmact) database
35. Ryoo Michael S, Matthies Larry (2013) First-person activity recognition: What are they doing to me? In: proceedings of the IEEE conference on computer vision and pattern recognition, p 2730–2737
36. Alletto Stefano, Serra Giuseppe, Calderara Simone, Solera Francesco, Cucchiara Rita (2014) From ego to nos-vision: detecting social relationships in first-person views. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops, p 580–585
37. Song Sibio, Chandrasekhar Vijay, Cheung Ngai-Man, Narayan Sanath, Li Liyuan, Lim Joo-Hwee (2014) Activity recognition in egocentric life-logging videos. In: Asian conference on computer vision, p 445–458. Springer
38. Damen Dima, Doughty Hazel, Farinella Giovanni Maria, Fidler Sanja, Furnari Antonino, Kazakos Evangelos (2018) Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In: European conference on computer vision (ECCV)

Soo-Han Kang received his B.S. and M.S degrees in computer science and engineering from Seoul National University of Science and Technology, Seoul, Korea, in 2019 and 2021, respectively. His research interests include machine learning and human-robot interaction.

Ji-Hyeong Han received her B.S. and Ph.D. degrees in electrical engineering from KAIST, Daejeon, Korea, in 2008 and 2015, respectively. From 2015 to 2017, she was a Senior Researcher with the Electronics and Telecommunications Research Institute in Daejeon, Korea. Since 2017, she has been with Seoul National University of Science and Technology, Seoul, Korea, where she is currently an Assistant Professor. Her research interests include machine learning, human-centered intelligent robotics, and human-robot interaction.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.