J Inf Process Syst, Vol.14, No.6, pp.1318~1330, December 2018
https://doi.org/10.3745/JIPS.02.0098

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Video Captioning with Visual and Semantic Features

Sujin Lee* and Incheol Kim**

Abstract

Video captioning refers to the process of extracting features from a video and generating video captions using the extracted features. This paper introduces a deep neural network model and its learning method for effective video captioning. In this study, visual features as well as semantic features, which effectively express the video, are also used. The visual features of the video are extracted using convolutional neural networks, such as C3D and ResNet, while the semantic features are extracted using a semantic feature extraction network proposed in this paper. Further, an attention-based caption generation network is proposed for effective generation of video captions using the extracted features. The performance and effectiveness of the proposed model is verified through various experiments using two large-scale video benchmarks such as the Microsoft Video Description (MSVD) and the Microsoft Research Video-To-Text (MSR-VTT).

Keywords

Attention-Based Caption Generation, Deep Neural Networks, Semantic Feature, Video Captioning

1. Introduction

The development of artificial intelligence in various fields, such as computer vision, natural language processing, and machine learning, has led to a growing interest in complex intelligence problems that require simultaneous processing of natural language and images. With the exponential rise in the generation of video data, led by the popularization of online video sharing platforms such as YouTube, Dailymotion, and Netflix, research interest in automatic analysis of video has grown. Typical video-based complex intelligence problems include video captioning and video question-answering. In this study, we attempt to add to the existing research on video captioning methods. Video captioning refers to the problem of generating natural language sentences that describe the input video, as shown in the example in Fig. 1. Video captioning technology can be applied to automatic video subtitle generation, video contents search, and video understanding.

This study proposes a deep neural network model, SeFLA, for video captioning. The proposed video captioning model uses not only visual features, but also semantic features to effectively represent a video. Visual features are extracted using the existing ResNet (residual network) [1] and C3D (3D ConvNet) [2] convolutional neural networks (CNN), while semantic features are extracted using a novel semantic feature network. Semantic features are divided into dynamic semantic features

Manuscript received November 29, 2017; first revision January 8, 2018; second revision March 13, 2018; accepted June 30, 2018.
Corresponding Author: Incheol Kim (kic@kyonggi.ac.kr)
* Dept. of Computer Science, Graduate School of Kyonggi University, Suwon, Korea (dltnwls9623@gmail.com)
**Dept. of Computer Science, Kyonggi University, Suwon, Korea (kic@kyonggi.ac.kr)

representing actions in the video and static semantic features representing objects, people, and background. These features are used as input features. To generate captions effectively based on these features, a selective attention caption generation network is proposed in this study. The selective attention caption generation network determines which semantic features among the input features to focus on in each timestep before generating captions. To analyze the performance of the proposed model, various experiments were performed using two large-scale video benchmarks, such as the Microsoft Video Description (MSVD) [3] and the Microsoft Research Video-To-Text (MSR-VTT) [4].

Fig. 1. Example of video captioning.

2. Related Works

Video captioning can be formulated as a problem of receiving a sequence of frames and outputting a sequence of words. Given this characteristic of video captioning, existing studies used an encoder-decoder framework, which had been mainly adopted in machine translation [5-7]. The encoder-decoder framework for video captioning consists of two parts: one part to extract features representing a video using an encoder, and the other part to output word sequences of the caption using a decoder. In particular, CNN models such as pre-trained very deep convolutional network (VGG) [8], ResNet, and C3D have been used as encoders, while recurrent neural network (RNN) has been mainly used as a decoder.

Moreover, studies applying soft attention to features have also been published [9,10]. Applying soft attention means that the video features extracted by the CNN model are not used as they are, but different weights are assigned to the contents of the video features. This makes it possible to use features that represent the input video more effectively, resulting in better caption generation performance. There are two types of soft attention: temporal attention and spatial attention.

Temporal attention indicates which frames in a frame sequence of the video to focus on, whereas spatial attention indicates which space to focus on in a frame. At present, in order to obtain effective input features, temporal attention or spatial attention is applied depending on the video features whenever words are generated for caption sentences.

Further, the use of semantic features in addition to visual features has been researched as a method to obtain more effective input features for caption generation [11-14]. As explained, semantic features refer to words representing actions, objects, people, background, etc., of videos. The RNN model for caption generation can understand the video more effectively by using semantic features consisting of direct expressions about the video. There are several methods that use semantic features; for example, using semantic features simply as input of a RNN for generating captions in each timestep, using semantic features as internal parameter weights of a RNN [11], applying attention to semantic features

[12], and embedding the extracted semantic features [13]. However, existing studies are limited in that they do not distinguish semantic features collected from the verbs and nouns of the caption dataset, and use a relatively simple long-short-term-memory (LSTM) model for generating captions using semantic features [11-13].

On the other hand, studies on RNN models for caption generation have also been studied. Existing studies mostly used the LSTM model, which is a type of RNN for caption generation, because compared to other RNN models, past LSTM enables improved memorization of information. Moreover, LSTM models of different frameworks were also researched [15,16]. In particular, Song et al. [15] used two hierarchical LSTM models for caption generation and applied temporal attention, which enables the determination of frames to be focused on from among the visual features extracted from ResNet between the two LSTMs.

3. Video Captioning Model

3.1 Model Overview

In this study, to realize effective video captioning, the use of semantic features as well as visual features representing videos is proposed. By applying them collectively, visual features extracted using a CNN represent the attributes of the input video including people, objects, background, actions, etc. To directly use high-level concepts inherent in visual features, the extracted words from visual features are used as semantic features. This enables the caption generation model to effectively understand the input video through the representative words describing the input video. In this study, the extracted semantic features are divided into dynamic and static semantic features, which are then used as additional features of the selective attention caption generation network.

Fig. 2. Examples of (a) dynamic and (b) static semantic features.

As shown in Fig. 2, dynamic semantic features express behaviors in the video, whereas static semantic features express objects, people, background, etc. In other words, the words corresponding to verbs in caption sentences are dynamic semantic features and nouns are static semantic features. The overall structure of the proposed model is shown in Fig. 3. First, visual features required to generate captions are extracted using pre-trained ResNet and C3D. These extracted visual features are input to the dynamic semantic network (DSN) and the static semantic network (SSN), which are introduced in Section 3.2. Then, the dynamic and static semantic features are extracted from each semantic network. These extracted semantic features are concatenated and input to the attention-based caption generation

network (CGN), which is introduced in Section 3.3, in each timestep. The visual features extracted through ResNet are used as input for the SSN and the LSTM which encodes the visual features. The final output of the encoding LSTM is provided for the initialization of the CGN. The CGN determines which semantic features to focus on in each timestep and calculates the probability distributions of words. Then, the captions are generated through the probability distributions of the output words.

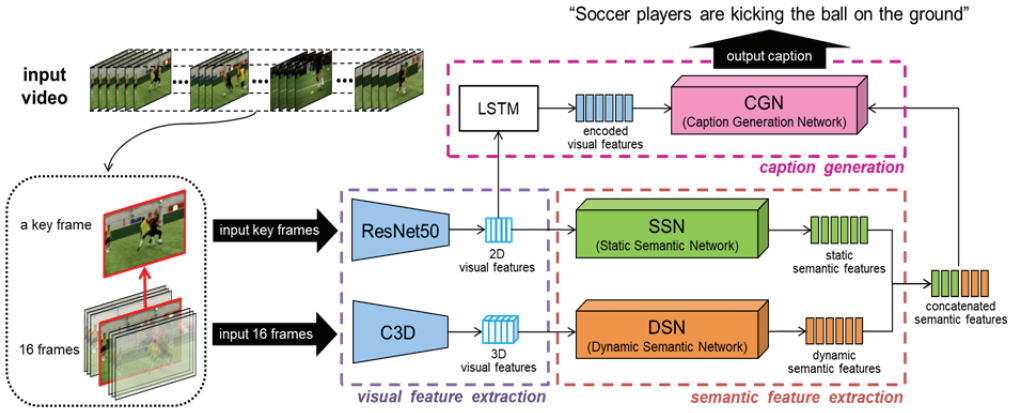

Fig. 3. Video captioning model.

3.2 Learning Semantic Features

To generate captions using semantic features, they must be obtained from the input video. These dynamic and static features have clear differences. Dynamic features are difficult to capture from one scene or frame of the video; they are extracted by observing the video for a certain time. On the other hand, static features such as objects, people, and background can be captured by observing one frame of the video. Therefore, the extracted semantic features are divided into dynamic and static semantic features in this study; this can be regarded as a multi-label classification problem. In particular, dynamic semantic features are extracted using visual features that effectively express the spatiotemporal characteristics of the video. On the other hand, static semantic features are extracted using visual features that effectively express the spatial characteristics of the video.

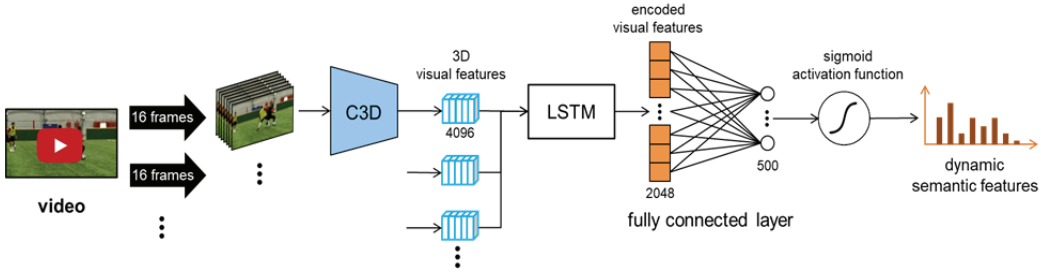

Fig. 4. The dynamic semantic network (DSN).

The proposed DSN is illustrated in Fig. 4. First, to capture visual features that effectively express the spatiotemporal characteristics of the video, visual features are extracted from every 16 frames of the video from a pre-trained C3D CNN, as shown in Eq. (1). Here, v_i denotes a single frame of the video,

n_v denotes the total frame count of the video, and $\frac{n_v}{16}$ denotes the total number of clips when the video is divided in 16 frame units. Then, the extracted visual features are encoded as shown in Eq. (2) by using the LSTM RNN model. Here, c_t denotes the visual features corresponding to one clip that is to be encoded in the current timestep (t), and h_{t-1} denotes the hidden state before LSTM.

$$c_i = C3D(v_{i:i+16}), i \in \left\{0, 1, \cdots, \frac{n_v}{16}\right\} \tag{1}$$

$$e = LSTM(c_t, h_{t-1}) \tag{2}$$

Next, from the encoded visual features, the probability distributions (p_d) of the dynamic semantic features are determined through the fully connected layer and the sigmoid activation function, as shown in Eq. (3); here, W_d denotes the learning weight and b_d denotes the bias.

$$p_d = sigmoid(W_d \cdot e + b_d) \tag{3}$$

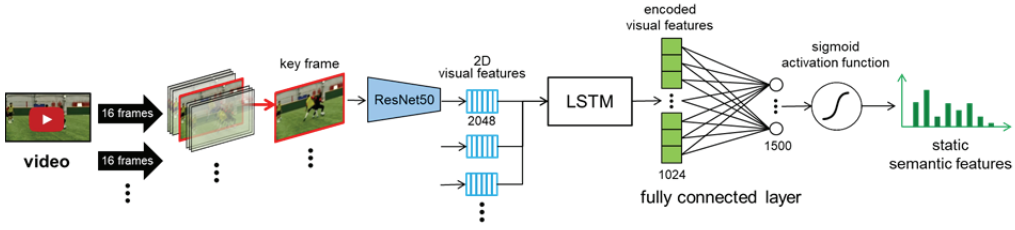

Fig. 5. The static semantic network (SSN).

The proposed SSN is illustrated in Fig. 5. First, to capture visual features that effectively express the spatial characteristics of the video, they are extracted from the pre-trained ResNet CNN. After the video is divided in 16 frame units, visual features (r_i) are extracted from the eight frames corresponding to the middle frame, as shown in Eq. (4), and encoded using the LSTM, as shown in Eq. (5). Then, the probability distributions (p_s) of the static semantic features are determined using the fully connected layer and the sigmoid activation function.

$$r_i = ResNet(v_{i\times 8+16}), i \in \left\{0, 1, \cdots, \frac{n_v}{16}\right\} \tag{4}$$

$$e = LSTM(r_t, h_{t-1}) \tag{5}$$

$$p_s = sigmoid(W_s \cdot e + b_s) \tag{6}$$

3.3 Attention-Based Caption Generation

In this study, an attention-based caption generation network is proposed for effective caption generation using semantic features. The proposed CGN is schematically illustrated in Fig. 6.

The CGN receives dynamic and static semantic features as input in each timestep and determines their probability distributions. These dynamic and static semantic features are concatenated and input to the attention layer. In general, when generating captions, focusing on the objects in the video is more effective when the currently generated word is a noun; on the other hand, focusing on the behaviors in the video is more effective if the currently generated word is a verb. To implement this selection, the

attention layer is used to determine which semantic features to be focused on in the current timestep. The attention layer calculates the semantic features by applying a weight (W_a) that indicates the semantic features to be focused on in the current timestep (t). These weighted semantic features (a_t) are calculated by using Eq. (7), where s_t denotes the input semantic features and b_a denotes the bias.

$$a_t = softmax(W_a \cdot s_t + b_a) \tag{7}$$

Next, the semantic features are input to the decoding LSTM. The decoding LSTM learns the sentence structures and outputs a state value indicating which words to generate in the current timestep, as shown in Eq. (8). The initial hidden state ($h_{t=0}$) of the decoding LSTM is initialized as the last hidden state of the encoding LSTM that encodes visual features.

$$h_t = LSTM(a_t, h_{t-1}) \tag{8}$$

The output of the decoding LSTM is input to the fully connected layer again. The fully connected layer calculates the probability distribution (p_t), which indicates the appropriate word in the current timestep (t), as shown in Eq. (9). Here, W_p denotes the learning weight, h_t denotes the input from the decoding LSTM, and b_p denotes the bias.

$$p_t = softmax(W_p \cdot h_t + b_p) \tag{9}$$

The attention for the input semantic features is calculated in each timestep and the probability distribution of each word is output through the decoding LSTM and the fully connected layer. Then this process is repeated, and the words are generated as captions from the first word to the last word <EOS> indicating the end of the sentence.

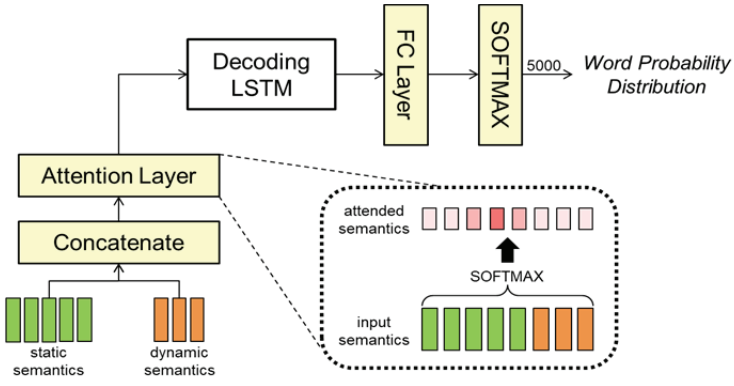

Fig. 6. The caption generation network (CGN).

4. Evaluation

4.1 Datasets

To train and evaluate the proposed caption generation model, two large-scale video benchmark datasets, the MSVD [3] and the MSR-VTT [4], were used. The MSVD data set consists of 1,970

YouTube video clips and around 80,000 caption sentences. The video clips are divided into 1,200, 100, and 670 videos and used as the training, validation, and test sets, respectively. On the other hand, the MSR-VTT data set consists of around 10,000 web video clips. The video clips are classified into 20 categories: music, people, gaming, sports/actions, news/events/politics, education, TV shows, movie/comedy, animation, vehicles/autos, how-to, travel, science/technology, animals/pets, kids/family, documentary, food/drink, cooking, beauty/fashion, advertisement. They are divided into 6,513, 497, and 2990 videos for training, validation, and test sets, respectively. Each video has around 20 natural language captions.

To learn the proposed semantic network, the MSVD video caption data set was used for learning the semantic features. First, the caption sentences of the MSVD data set were separated into nouns and verbs using the part-of-speech (POS) function of the natural language toolkit (NLTK). Then, the plural forms of nouns and the past, present-continuous, and other tenses of verbs were transformed into basic forms using the lemmatize function of NLTK. Next, the label data of dynamic semantic features were composed by selecting the first 500 verbs in order (descending) of their frequency of appearance from among the separated verbs; similarly, the label data of static semantic features comprised of first 1,500 nouns in order of their frequency of appearance from among the separated nouns. If a specific word in the label data of dynamic semantic features was included in the video captions, the video for this verb was labeled as 1; otherwise, it was labeled as 0. In this way, a dynamic data set was generated; similarly, a static data set was generated. In this data set, each video is associated with approximately seven nouns and three verbs. The semantic feature data sets consisted of 1,200, 100, and 670 videos for training, validation, and test sets, respectively, as with the MSVD caption data sets.

4.2 Model Training

The proposed model was implemented using Keras, which is a Python deep learning library, in the Ubuntu 14.04 LTS environment. The videos to be used as input were sampled uniformly such that each video was composed of 40 clips of 16 frames. The semantic network was trained using Adam as the model optimization algorithm and using the binary cross-entropy expressed in Eq. (10) as loss function, where y denotes the correct answer and $\tilde{y}$ denotes the predicted value.

$$L_{binary} = -[y \log \tilde{y} + (1-y) \log(1-\tilde{y})] \tag{10}$$

After the completion of training of the semantic feature extraction network, the semantic features are extracted from all videos in the caption data set in advance. Then, the extracted features are used as input to the CGN. The RMSprop algorithm was used as the model optimization algorithm for the CGN, The binary cross-entropy expressed in Eq. (11) was used as the loss function.

$$L_{categorical} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y_i})^2 \tag{11}$$

For semantic feature extraction model, the batch size and the epoch were set at 32 and 500, respectively. On the other hand, for the caption generation model, the batch size, and the epoch were set at 25 and 50, respectively.

4.3 Quantitative Analysis

In the first experiment, the performance of proposed semantic feature extraction networks was evaluated. The accuracy of each semantic feature extraction network was calculated using the mean square error (MSE) as shown in Eq. (12), where n denotes the output dimension, y_i represents the correct value, and $\tilde{y}_i$ denotes the predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{12}$$

Table 1 shows the results of the performance evaluation for each network. In Table 1, DSN refers to dynamic semantic network and SSN refers to static semantic network. Both networks achieved high accuracies of semantic feature extraction.

Table 1. Performance comparison between two semantic networks on the MSVD dataset

Network	Accuracy (%)	
	Val	Test
DSN	99.42	99.43
SSN	99.61	99.64

In the second experiment, the effects of each semantic feature on the caption generation performance were examined. The proposed selective attention caption generation network was used as the CGN, while the input features were varied. BLEU@N [17] and CIDEr-D [18], which are general caption generation evaluation metrics, were used to measure the performance of the CGN. Specifically, BLEU scores measure the fraction of n-grams that are in common between hypothesis and a reference or a set of references. CIDEr is consensus based evaluation protocol which rewards a sentence for being similar to the majority of human written descriptions.

Each evaluation metric was calculated using the code provided by the Microsoft COCO evaluation server. Table 2 shows the results of experiments conducted on the MSVD dataset. In Table 2, CGN refers to the generation of captions using only visual features without using semantic features; DSN+CGN refers to the use of dynamic semantic features only; SSN+CGN refers to the use of static semantic features only; and DSN+SSN+CGN refers to the use of both dynamic and static semantic features.

Table 2. Performance comparison among different feature models on the MSVD dataset

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr
CGN	66.1	47.8	37.1	26.5	26.4
DSN+CGN	76.0	58.1	45.7	35.8	50.0
SSN+CGN	78.8	63.4	51.4	41.4	77.8
DSN+SSN+CGN	84.8	70.8	60.0	50.0	94.3

Table 3. Performance comparison among different models on the MSR-VTT dataset

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr
CGN	68.5	44.1	29.3	14.7	7.9
DSN+CGN	68.2	43.6	28.6	16.8	16.9
SSN+CGN	70.7	48.6	32.8	21.6	34.1
DSN+SSN+CGN	77.7	62.0	50.4	41.8	60.1

Table 2 lists evaluation metrics for each model type. It can be seen that the semantic features-based models exhibited better performances. In particular, the model using static semantic features only yielded better performance than the model using dynamic semantic features only. This may be attributed to the fact that the dynamic semantic features represent only the actions in the video, whereas the static semantic features represent people, objects, and background; these features represent the video more accurately. The model using both semantic features yielded the best performance, which suggests that the two types of semantic features contributed independently to the caption generation performance.

Table 3 shows the results of experiments conducted on the MSR-VTT dataset. Similar to Table 2, it can be seen that the model using both semantic features yielded the best performance. In particular, Table 3 shows that static semantic features contribute more to the improvement of the caption generation performance than the dynamic semantic features.

In the third experiment, the performance of SeFLA, which is the caption generation model proposed in this study, was compared with that of other state-of-the-art models (SCN, LSTM-TSA, hLSTMat). Both SCN and LSTM-TSA use semantic features as well as visual features. However, unlike our SeFLA, they are limited in that they do not distinguish the dynamic semantic features from the static semantic features. Moreover, they use a relatively simple LSTM model for generating captions. On the other hand, hLSTMat uses an attention-based hierarchical LSTM for caption generation. However, unlike SCN, LSTM-TSA, and our SeFLA, the model is limited in that it uses only visual features, but not semantic features. Table 4 shows the performance comparison results between SeFLA and other state-of-the-art models on the MSVD dataset.

Table 4. Performance comparison with other state-of-art models on the MSVD dataset

Models	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr
SCN [11]	-	-	-	51.1	77.7
LSTM-TSA [12]	82.8	72.0	62.8	52.8	74.0
hLSTMat [15]	82.9	72.2	63.0	53.0	73.8
SeFLA	84.8	70.8	60.0	50.0	94.3

SeFLA yielded 84.8 in BLEU@1 and 94.3 in CIDEr, higher by 1.9% and 16.6% compared to the existing studies. The performance scores for BLEU@2, 3, and 4 were lower than those of other studies. This indicates that the proposed SeFLA can capture single words correctly, but it yields poor performance when capturing multiple words consecutively. This implies that SeFLA can generate nouns and verbs in captions with the help of semantic features; however, it cannot accurately generate prepositions and postpositions, which are necessary for sentence construction. The limitation in performance can be attributed to the fact that the lack of sufficient training data of sentence structure for training of LSTMs in the caption generation network. In general, the proposed SeFLA generates accurate captions by effectively applying semantic features.

4.4 Qualitative Analysis

On the MSVD dataset, we conducted qualitative evaluation of the proposed SeFLA model, which makes use of semantic features. Fig. 7 shows some correct captions generated by the proposed SeFLA model. In Fig. 7, GT denotes the ground-truth caption of the video. In these cases, the SeFLA model not only extracted the relevant semantic features successfully, but also composed the correct captions with

these semantic features. On the other hand, Fig. 8 illustrates some cases generating incorrect captions. Most of them are the cases where the SeFLA model extracted successfully the relevant semantic features, but unfortunately generated incorrect captions with these semantic features. The results show the power of our semantic feature networks, while they suggest further improvement of the precision of the caption generation network.

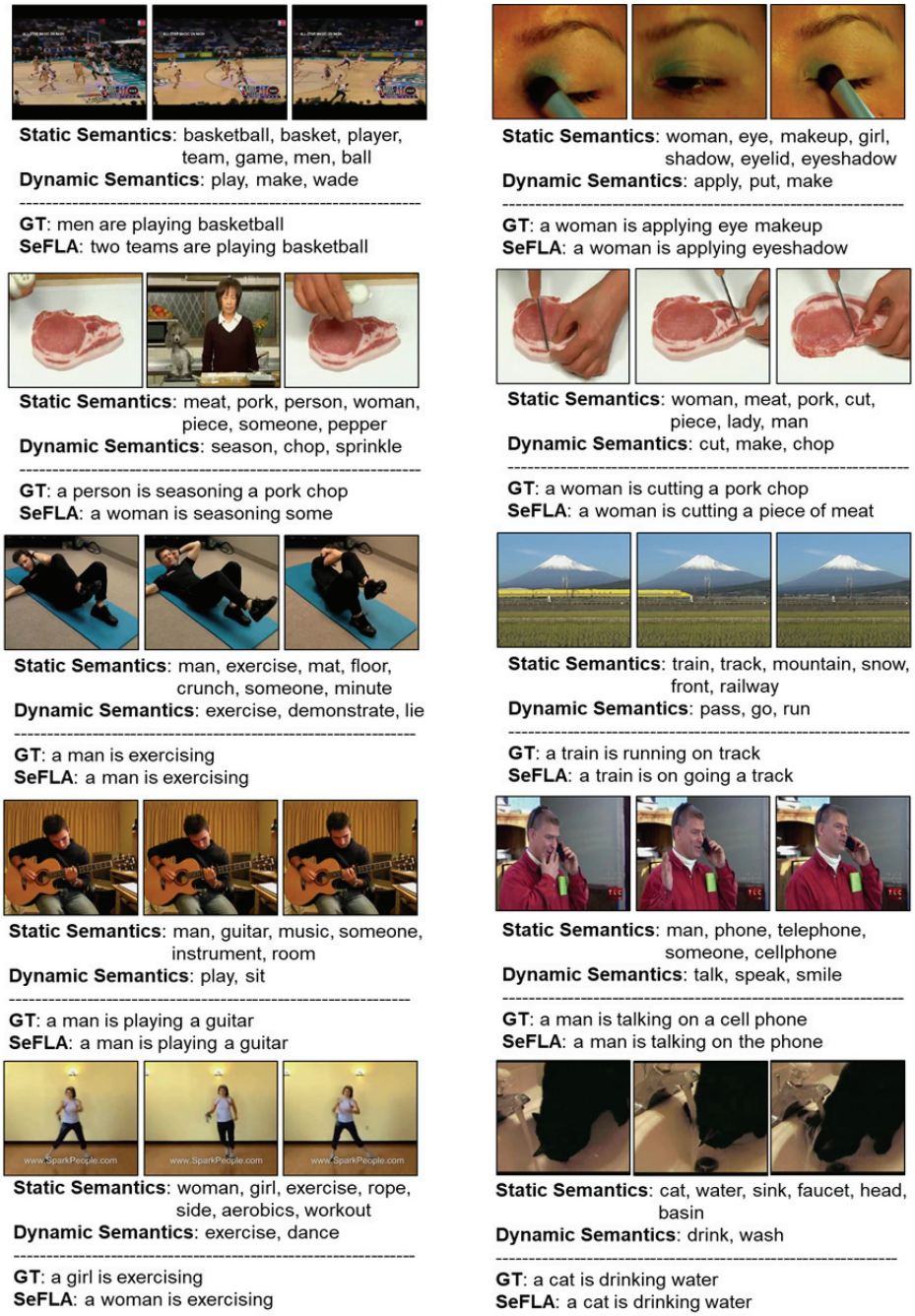

Fig. 7. Qualitative results on the MSVD dataset: correct captions with relevant semantic features.

Fig. 8. Qualitative results on the MSVD dataset: incorrect captions with relevant semantic features.

5. Conclusion

In this study, an effective deep neural network model was proposed for video captioning. The proposed caption generation model uses semantic features from two semantic feature extraction networks as well as visual features extracted using a CNN from the input video. Our model uses two distinct types of semantic features. While dynamic semantic features represent actions included in the video, static semantic features represent objects, people, or places. Furthermore, our model employs an attention-based caption generation sub-network, which focuses attention on semantic features in a video for more effective caption generation. Through various experiments using two large-scale video benchmark datasets such as the MSVD and the MSR-VTT, the proposed SeFLA model showed better performance than the state-of-art models. In the future, we plan to conduct more experiments on other benchmark datasets to check the scalability of the proposed SeFLA model. We also consider extending the SeFLA model to generate dense captions of a video in order to help understanding its content.

Acknowledgement

This research was supported by the Ministry of Science and ICT, Korea, under the Information Technology Research Center support program (No. IITP-2017-0-01642) supervised by the Institute for Information & Communications Technology Promotion (IITP).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4489-4497.

[3] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 2712-2719.

[4] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: a large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 5288-5296.

[5] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence: video to text," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4534-4542.

[6] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 4594-4602.

[7] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4507-4515.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[9] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 1141-1150.

[10] Y. Pan, T. Yao, H. Li, and T. Mei, "Video Captioning with Transferred Semantic Attributes," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 984-992.

[11] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 3261-3269.

[12] F. Nian, T. Li, Y. Wang, X. Wu, B. Ni, and C. Xu, "Learning explicit video attributes from mid-level representation for video captioning," *Journal of Computer Vision and Image Understanding*, vol. 163, pp. 126-138, 2017.

[13] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, and H. T. Shen, "Hierarchical LSTM with adjusted temporal attention for video captioning," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, Melbourne, Australia, 2017, pp. 2737-2743.

[14] A. A. Liu, N. Xu, Y. Wong, J. Li, Y. T. Su, and M. Kankanhalli, "Hierarchical & multimodal video captioning: discovering and transferring multimodal knowledge for vision to language," *Journal of Computer Vision and Image Understanding*, vol. 163, pp. 113-125, 2017.

[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 311-318.

[16] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 4566-4575.

Sujin Lee https://orcid.org/0000-0003-0634-6240

She received B.S. degree in Computer Science from Kyonggi University in 2017. She is currently a M.S. student of Department of Computer Science, Kyonggi University, Korea. Her current research interests include machine learning, computer vision, and intelligent robotic systems.

Incheol Kim https://orcid.org/0000-002-5754-133X

He received the M.S. and Ph.D. degrees in Computer Science from the Seoul National University, Korea, in 1987 and 1995, respectively. He is currently a Professor of the Department of Computer Science, Kyonggi University, Korea. His current research interests include machine learning, knowledge representation and reasoning, task and motion planning, computer vision, and intelligent robotic systems.