INTERNATIONAL JOURNAL OF ENGINEERING BUSINESS MANAGEMENT

International Journal of Engineering Business Management
Volume 14: 1–11

DOI: 10.1177/18479790221078130
journals.sagepub.com/home/enb
SAGE

A core region captioning framework for automatic video understanding in story video contents

Hyesun Suh[1], Jiyeon Kim[2], Jinsoo So[3] and Jongjin Jung[1]

Abstract

Due to the rapid increase in images and image data, research examining the visual analysis of such unstructured data has recently come to be actively conducted. One of the representative image caption models the DenseCap model extracts various regions in an image and generates region-level captions. However, since the existing DenseCap model does not consider priority for region captions, it is difficult to identify relatively significant region captions that best describe the image. There has also been a lack of research into captioning focusing on the core areas for story content, such as images in movies and dramas. In this study, we propose a new image captioning framework based on DenseCap that aims to promote the understanding of movies in particular. In addition, we design and implement a module for identifying characters so that the character information can be used in caption detection and caption improvement in core areas. We also propose a core area caption detection algorithm that considers the variables affecting the area caption importance. Finally, a performance evaluation is conducted to determine the accuracy of the character identification module, and the effectiveness of the proposed algorithm is demonstrated by visually comparing it with the existing DenseCap model.

Keywords

Core region detection algorithm, image captioning models, video scene analysis, proposed algorithm, DenseCap model

Date received: 10 November 2021; accepted: 18 January 2022

Introduction

Image and video data generated from smartphones, high-definition cameras, CCTV, drones, etc. are increasing exponentially, and video contents such as Netflix, YouTube, and Internet TV are expected to account for 82% of all IP traffic in 2021. There is a growing demand for technology that can visually analyze such unstructured data to derive meaningful information. In the field of deep learning, research is actively being conducted in image classification to predict multiple labeled images to analyze images and videos, object detection to predict and label objects in images, and captioning to generate descriptions of images in natural language form.[1–5]

Usually, captions are generated for a whole scene in the image. Image captioning methods can use either simple encoder-decoder architecture or compositional architecture. In encoder-decoder architecture-based methods, global image features are extracted from the activations of CNN as encoder and then fed them into an LSTM as decoder to generate sentence. Oriol et al proposed a method called

[1]Department of Computer Science Engineering, Daejin University, Pocheon-si, Korea
[2]Department of Creative Future Talents, Daejin University, Pocheon-si, Korea
[3]BigData Team, Information&Communication Division, Goyang-si, Korea

Corresponding author:
Jiyeon Kim, Department of Creative Future Talents, Daejin University, 1007 Hoguk-ro, Pocheon-si 11159, Korea.
Email: jykim629@daejin.ac.kr

Neural Image Caption Generator (NIC), but this method had the problem of vanishing gradient problems caused by learning process of LSTM.[6] To solve this problem,[7] an extended LSM called guided LSTM(gLSTM) has been proposed. Junhua et al.[8] proposed a special image captioning method which generated caption for a specific object or region.[8] In compositional architecture-based methods, semantic concepts of image are extracted CNN. And then a language model is used to generate a set of candidate captions and deep multimodal similarity model re-ranks them to generate the final caption.[9–11]

Because the image captioning model only generates a single caption for the entire image, it is difficult for this model to sufficiently describe the regions and objects in the image in details. To solve this problem, recent studies have proposed the use of dense captioning models to extract various regions in an image and generate region-level captions through the integration of an object detection model and an image captioning model to generate richer captions.[12–18] This dense captioning is based on the Visual Genome[19] region caption data set based on MS COCO and the YFCC100M image data set. However, although this DenseCap model generates various region captions, it does not consider varying levels of importance for these region captions, so it leads to many unnecessary area captions, this making it difficult to identify area captions that are relatively important in interpreting images. Furthermore, captioning research focusing on the core region for the story content of movies and dramas is still very insufficient.

Therefore, in this study, we propose an image captioning framework centered on a core region for the automatic understanding of movies, which is a type of video content that is driven by storytelling. The proposed framework is based on the DenseCap model,[20] and the process consist of (1) character identification through a character identification module, (2) the operation of the proposed caption detection algorithm in core regions, and (3) a post-processor procedure. First, a character face learning dataset is created using the HoG algorithm[21] and DBSCAN so that information on the characters in the movie can be used for caption detection and caption improvement in core regions. Thus, we design and implement a character identification module that identifies characters in movie scenes using the EfficientNet model.[22] Next, we propose a core region detection algorithm which considers four variables that affect the importance of region captions: "region box confidence score," "region area," "distance between the region and the center of the image," and "object type." Finally, the post-processor unifies region captions for the same object into a single caption and improves the region captions into a form suitable for a movie. The results show that the proposed framework effectively detects important region captions of movie scenes, and it therefore provides a lot of information with fewer region captions than the existing DenseCap captioning model.

Proposed algorithm

Algorithm construction

Recently, some researchers have focused on dense captioning which can generate captions by regions of objects in a scene. One captioning for whole scene is so subjective but dense captioning is more objective than one captioning. Justin et al. proposed a dense captioning method which is called DenseCap.[23] This method has a Fully Convolutional Localization Network (FCLN) which is composed of a convolutional network, a dense localization layer, and an LSTM language model. It localizes all the prominent regions of an image and generates captions for the regions. To do this localization work, it uses spatial soft attention and bilinear interpolation instead of ROI pooling in Faster R-CNN. It uses Visual Genome dataset to produce region captions. It also uses LSTM with region codes as a language model. But, there are some challenges in dense captioning. An object may have many overlapped regions because regions can be dense. Linjie et al proposed another model pipeline which is based on joint inference and context fusion.[24] However, as mentioned in the introduction, DenseCap model does not consider varying levels of importance for these region captions, so it leads to many unnecessary area captions, thus making it difficult to identify area captions that are relatively important in interpreting images.

Figure.1 shows the framework implementation procedure proposed in this study. First, a frame-by-frame scene image set is generated by extracting one frame per second from a movie. The scene image is the subject of image captioning, and a preprocessing process is performed to convert the width of the image to 720 pixels. Then, the DenseCap model generates 1000 region proposals for each scene image, and it sets the threshold of the initial Non Maximum Suppression to 0.7 and the final NMS threshold to 0.3 in order to reduce unnecessary overlapping of regions. As a result, an average of 105 area captions is generated for one scene image. (1) Key region caption detection algorithm of Figure 1 identifies the relatively important region caption for explaining the scene by considering the four variables ("confidence score," "region area," "distance between the area and the center of the scene," and "object type") presented for a large number of region captions extracted through the existing DenseCap model. After that in (2) post-processing is performed, such as integrating the region and caption for the same object into one and converting a word referring to a person into a character name. At this time, (3) character recognition module is used to classify the main character, supporting role, and extras in (1) when the "object

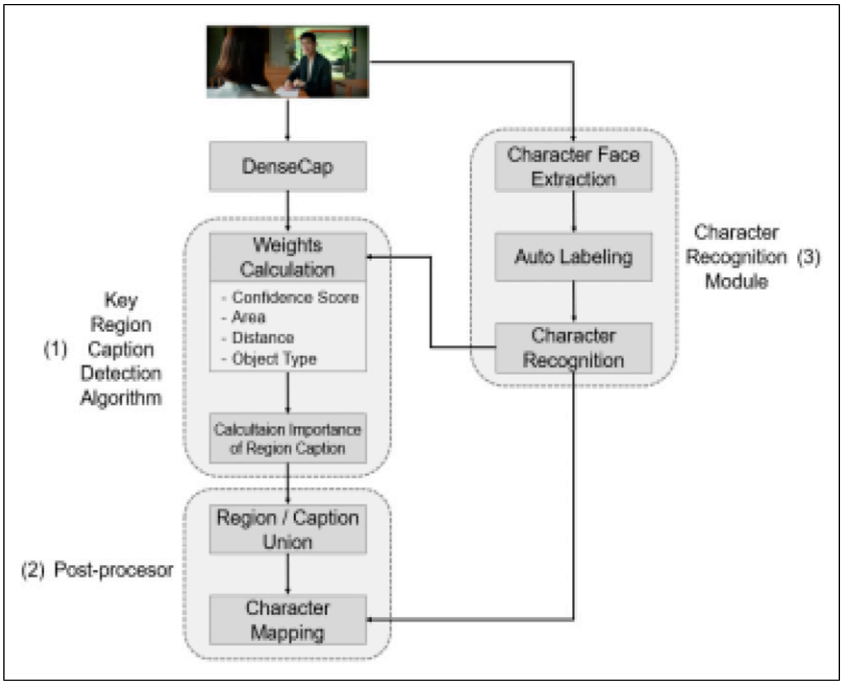

Figure 1. Proposed algorithm process.

type" is a person, and to convert the word referring to a person with a character name in (2). This module performs a series of process from character for data set, labeling, and face prediction.

Character identification module

The character identification module proposed in this study predicts the name and face box of each character appearing in a scene image using a face recognition model. The character's name and face box predicted by this character identification module are used to give importance to that movie character when the "object type" is a person, and it is also used when the post-processor converts the word that refers to the name of the region caption into a character name.

To generate a training data set for such a model, face images are collected by cropping faces as recognized by the Haar-Casade and HoG algorithms which are representative feature extraction techniques from a set of scene images. In the empirical analysis, the accuracy of face recognition by HoG was found to be higher than that by Haar-Casade, so in this paper, the HoG algorithm was ultimately used to collect face images.

Since there is no label required for supervised learning in the collected face images, similar face images are clustered by performing DBSCAN, a density clustering technique, on the HoG feature vectors of the face images to provide labels. When similar face images are clustered, the name of the character represented by the clusters is given as a label. Then, we proceed to directly reclassify the wrongly clustered face images while excluding extra face images. Finally, 30 augmented images are secured per each original image through augmentation that transforms the face image by cropping, rotating, flipping, translating, and resizing.

This study compared the performance in terms of accuracy of three models, VGG-16,[25] ResNet101, and EfficientNet-B5, which are CNN models that have been used for the previously generated face image training dataset. As a result of training in the same experimental environment and conditions, the EfficientNet model, which shows the best performance with a small number of parameters, was selected as the face recognition model used in this study.

Core region caption detection algorithm

This chapter deals with the algorithm that gives priority to region captions created through DenseCap. Here, we explain the details of the variables that affect the importance of the region caption as well as the method used to calculate the importance.

Region box confidence score. The confidence score of the domain box is a confidence index for the domain box having a larger value as the probability that an object existing in the domain box increases, or the domain box and the ground

true box match. Therefore, it is assumed that the higher the confidence score of the region box, the higher the probability that the region box accurately contains the object. Consequently, the higher the confidence score of the region box, the more important the region. The confidence score of the region box is calculated using equation (1). Confidence Score$_i$ means the confidence score of the ith region, and Pr(Object) means the probability that an object exists in the region box. IoU(Truth,B$_i$) is the ratio of the area where the correct answer box Truth and the i th region box B$_i$ overlap.

$$Confidence\ Score_i = Pr(Object) \times IoU(Truth, B_i) \quad (1)$$

Region area. Generally, to emphasize an important object in a movie, close-up photography techniques are used to increase the region occupied by that object in the entire scene. This assumes that the larger the region an object occupies in the screen region, the greater the probability that the object is important. The area of the region is calculated using equation (2). Here, Area$_i$ denotes the area of the i th region, and B$_{i,W}$ and B$_{i,H}$ respectively denote the width and height of the ith region box.

$$Area_i = B_{i,w} \times B_{i,H} \quad (2)$$

Distance between the region and the scene center. In addition, the more important the object in the scene, the more common it is for that object to be shot in the center of the frame. Therefore, it is assumed that the closer the object is to the center of the scene, the more likely it is to be important. In this study, the distance between the region and the center of the scene was calculated using the Euclidean distance measurement method as shown in the following equation, and the reciprocal was taken so that the shorter the distance, the larger the value. Here, S$_{centerX}$ and S$_{centerY}$ respectively refer to the X coordinate and Y coordinate of the center of the scene, and B$_{i,centerX}$ and B$_{i,centerY}$ respectively refer to the X coordinate and Y coordinate of the center of the ith region box. The i th region distance ultimately obtained in this way is called Distance$_i$.

$$Distance_i = \frac{1}{\sqrt{(S_{centerX} - B_{i,centerX})^2 + (S_{centerY} - B_{i,centerY})^2}} \quad (3)$$

Object type. In this paper, the noun extracted from the caption is defined as an object. Each object is categorized as a person, an object, or a background and a person can be further categorized as a main actor, a supporting role, or an extra according to the weight of the person. Assuming that the importance differs depending on the type of object, the importance of each object type is defined as presented in Table 1, which reflects expert interviews with three professors of theater and film at the university.

In this study, as shown in Figure 2, if more than 70% of the region box contains the face box of the character predicted in *Algorithm construction*, and the type of object represented by the region is a person, then that object is judged to be the predicted character, and this character is given importance. If the type of object is a person, but (1) the face of the person is not clear, (2) the face is an extra that cannot be identified, or (3) the proportion of the face box included in the region box is less than 70%, the object type is classified as an extra. If the type of object is not a person, it is classified as an object or background.

Importance calculation. Since the previously calculated "area box confidence score," "region area," "distance between region and scene center," and " object type" are variables with different units, the magnitude of their influence on the region caption importance is different. Therefore, standardized scaling is applied to each variable to convert it to a standard normal distribution. The scaled variables are denoted by Confidence Score′, Area′, Distance′, and Object Type′. The importance of the ith region caption is given by arranging the sum of the variables of the i th region that have been scaled in descending order, as shown in the following equation

$$Importance_i = Confidence\ \ Score'_i + Area'_i + Distance'_i + Object\ \ Type'_i \quad (4)$$

Post-processor

In the post-processing step, to improve the relevance of the caption and the upper area extracted through the proposal framework, the area and caption representing the same

Table 1. Importance by object type.

	Main actor	Supporting	Extra	Object	Background
Importance	3	2	1	1	0

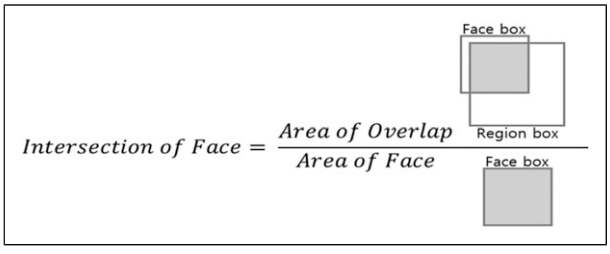

Figure 2. Percentage at which the face box is included in the area box.

object are integrated, and the word referring to the person in the area caption is converted into a character name. Post-processing proceeds as follows.

1) If two regions represent the same object and the IoU is 0.2 or higher, the two regions are judged as region captions for the same object, and the region box and caption are integrated.
2) If the object represented by the region is a person and the region box contains more than 70% of the face box of the character predicted by the character identification module, then the word representing that person in the caption is converted into a character name in the manner described in Table 2. Articles such as "a" and "the" in front of character names are removed.

Experiment and result

Empirical data and experiment environment

In this study, an experiment was conducted with the movie "Parasite" by director Bong Joon-ho selected as the target of analysis. The movie used for the analysis is a 1920x804 video, and 7922 scene images were collected in total by extracting one frame per second. Then, a preprocessing process was performed to adjust the 1920x804 image to the dimensions of 720x301. The characters in "Parasite" to be identified in the character identification module were defined as a total of 10 people using movie metadata: 6 main actors and four supporting actors. To generate a training data set for the face recognition model, 2299 face images in total were collected by cropping the faces recognized by the HoG algorithm from the scene image set. Then, 489 extra face images were excluded from the DBSCAN results, and 529 images that had been incorrectly clustered were directly reclassified. Finally, 1810 face images were collected in total. In order to obtain additional facial images, augmentation was performed, and the remaining 54,300 face images were used as the training data set. The procedure of the character identification module is illustrated in Figure 3.

Regarding the hardware and software environment used to conduct the experiment, Intel(R) Xeon(R) Gold 5120 was used as the CPU, and the VGA was composed of two Tesla V100 SXM2 32GB units. Further, Ubuntu 16.04 LTS was used as the operating system, and CUDA 10.1 and CUDNN 7.5.1 were installed. The proposed framework was implemented in Torch7 and Tensorflow 2.1 environments using Lua and Python.

Empirical experiment

The performance comparison in terms of accuracy of the CNN models VGG-16, ResNet101, and EfficientNet33 models in the face image training dataset under the same experimental environment and conditions showed the following results: the VGG-16 model achieved an accuracy of 85.15%, the ResNet101 model achieved an accuracy of 91.61%, and the EfficientNet model showed an accuracy

Table 2. How to convert region caption characters.

	Number of people represented by region captions		
Predicted character within the region box	1 person	2 people	Remainder
A	A	A and the other	A and the other
A, B	A or B	A and B	A and B and the other
3 or more	A or B or	A or B or …	A and B and … and the other

Figure 3. Character identification module process.

of 99.85%. Therefore, as the EfficientNet model has the highest accuracy, it was selected for use as the face recognition model.

Figures 4–9 show the region captions of a specific scene created in the existing DenseCap model as well as the results of the top five region captions to which the core region caption detection algorithm is applied. Figures 4–8 compare the DenseCap model that does not consider the importance of region captions and the identification of region captions according to the criteria of four variables that affect the importance of region captions proposed in this study.

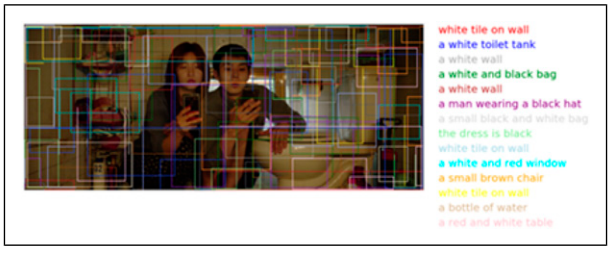

Figure 4. DenseCap model.

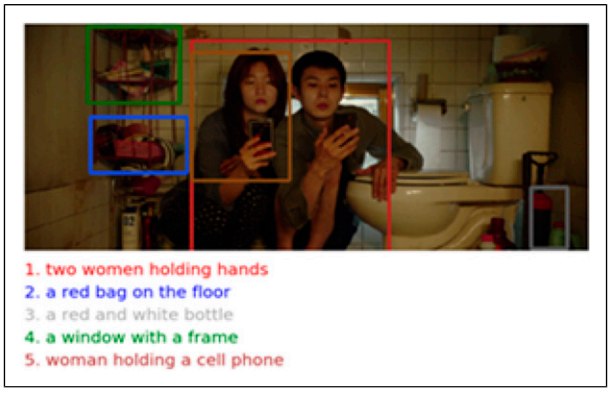

Figure 5. Proposed algorithm [Case1: Top five domain captions when considering domain box confidence scores].

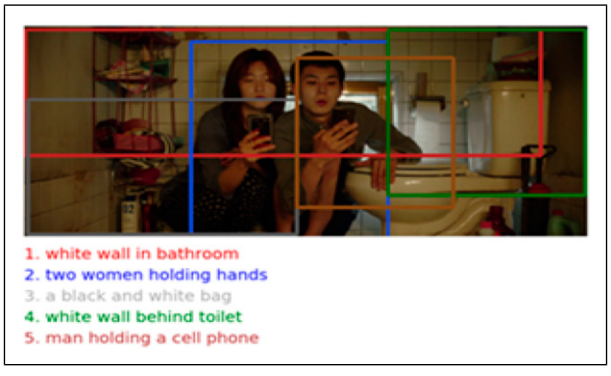

Figure 6. Proposed algorithm [Case2: Top five domain captions when considering region area].

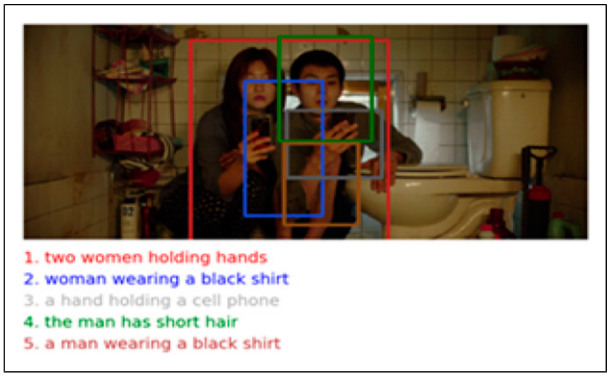

Figure 7. Proposed algorithm [Case3: Top five domain captions when considering the distance between the region].

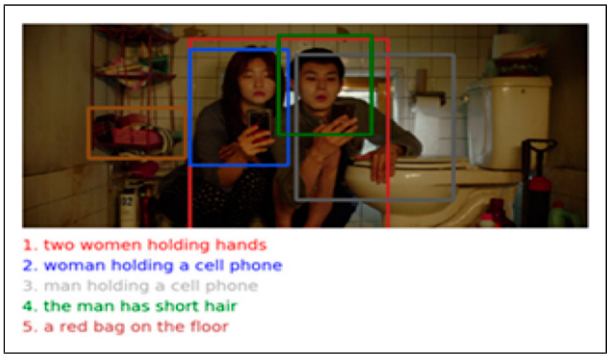

Figure 8. Proposed algorithm [Case 4: Top five domain captions with high importance by object type].

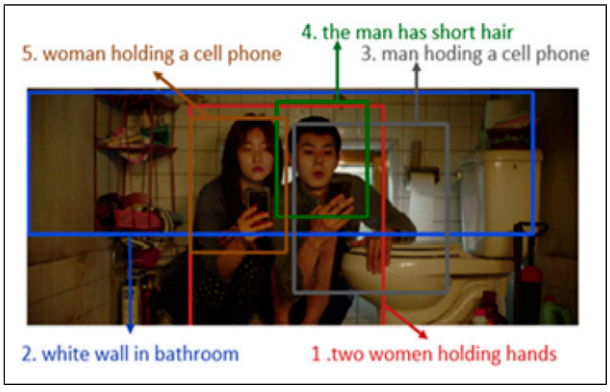

Figure 9. Proposed algorithm [Core region detection algorithm].

Figure 9 compares the degree of identification when four variables are considered at the same time. That is, we want to compare which models are good at identifying important area captions in the scene.

The results of the top five domain captions selected from each of "region box confidence score," "region area," "distance of between region and center of the scene," and "object type" which were considered to be variables

affecting the importance of domain captions in the proposed algorithm are also presented.

The DenseCap model has an average of 105 area captions for each scene. When only the " region box confidence score" of Case 1 in Figure 5 is considered, the region box accurately contains various objects appearing in the scene, but it can be seen that a large number of region captions that are relatively insignificant in the scene also appear, such as " a red bag on the floor," " a red and white bottle," " a window with a frame," etc. When only the " region area" of Case 2 in Figure 6 is considered, the objects that occupy a large area in the scene are mainly extracted, and the region captions such as "white wall in bathroom," "two women holding hands," and "man holding a cell phone" that are important in the scene are extracted as well. However, region captions for surrounding objects such as "a black and white bag" are also extracted. Next, in the region caption where the "distance between the region and the center of the scene" of Case 3 in Figure 7 was close, the region caption focused on the two figures being focused on in the scene, relatively insignificant details were shown as region captions as well, such as "a man wearing a black shirt" and "the man has short hair." When only the "object type" of Case 4 in Figure 8 was considered, the region captions including the faces of two people with high importance were extracted first, and these were followed by the appearance of captions describing objects. This can extract detailed region captions for the face, such as "the man has short hair," and extracting region captions for insignificant such as "a red bag on the floor" for objects or backgrounds of low importance. Lastly, Figure 9, the region captions extracted by the core region caption algorithm that simultaneously considers these four variables are "two women holding hands," "white wall in bathroom," "man holding a cell phone," "the man has short hair," and "woman holding a cell phone." Rather than considering each variable individually, it can be seen that the core region caption detection algorithm extracts region captions that better explain the scene in which two people are holding their cell phones in a bathroom.

Because the existing DenseCap model did not consider various elements, such as the character information and domain of the movie, it could not identify the relative importance of area captions in terms of understanding the movie scene. Meanwhile, the proposed algorithm identifies the captions of important regions of a movie scene better than the existing DenseCap model.

Figures 10–12 show the DenseCap model results, the top five region caption results by proposed model, and the results after post-processing. Figure 12 shows the results of extracting the captions of the top five regions according to the proposed framework and performing post-processing. It can be seen that the post-processing results are clearer. The top five region captions extracted by the core region detection algorithm are "two men standing in a room," "a woman wearing a white shirt," "woman with long hair," "man in black shirt," and "man with short brown hair." Among the region captions, the IoU values of the fourth and fifth area boxes are both 0.2 or more, and the object represents the person "man," so the region box and the caption are integrated. Similarly, the second and third region captions unite the region box and caption, as the object represents the character "woman." Further, since the "Yeon Kyo" character face box predicted by the character prediction module is included in more than 70% of the area box, "woman" in the region caption was converted to "Yeon Kyo," and the preceding article was removed. The first region caption is "two men standing in a room,"

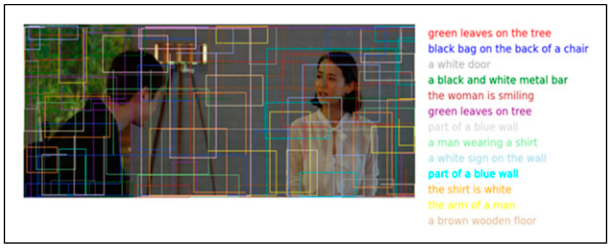

Figure 10. Existing DenseCap results.

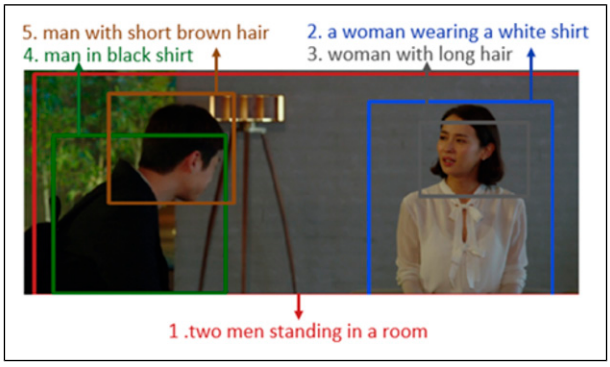

Figure 11. Top five region captions.

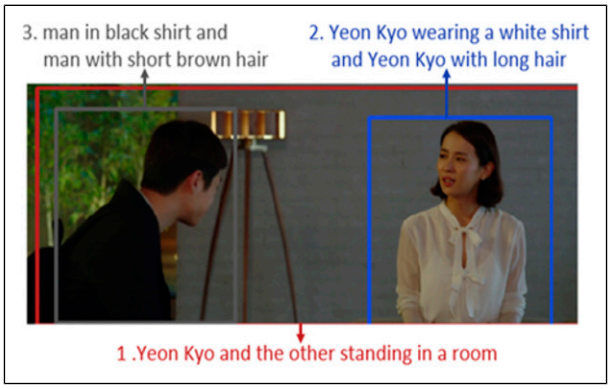

Figure 12. Post-processing results in the proposed framework.

representing two people, and since the character identification module in the region box predicted one character to be "Yeon Kyo," "two men" is converted to "Yeon Kyo and the other." The final converted area caption is reduced to three, and it can be seen that the area caption reflects the movie character information well.

Figure 13 compares the region caption generated by the existing dense captioning models, DenseCap and Dense Relational Captioning,[13] with a part of the region caption generated by the proposed framework. For this image, DenseCap generated 95 region captions and Dense Relational Captioning generated 8930 region captions. By contrast, the proposed framework generated 57 region captions by integrating the region captions for the same object. Among the region captions generated by DenseCap and Dense Relational Captioning, there are many region captions that are relatively insignificant for understanding a movie scene. Because Dense Relational Captioning which focuses on the relationships between objects also generates captions for pairs of objects with low relevance, it generates false region captions at a higher rate than DenseCap. In addition, in existing dense captioning models, it is difficult to specify which character is being described by a region caption that refers to a person using words such as "man," "woman," "people," "boy," and "girl." On the other hand, in the region caption generated through the proposal framework, when the person represented by the caption is identified, the character name is reflected in the caption, thus making it possible to know which character the caption describes.

In conclusion, the proposed framework significantly reduces the number of region captions and extracts region captions that are relatively important for understanding the scene of a movie. In addition, it shows improved results compared to the existing dense captioning model by using

DenseCap (N=95)	Dense Relational Captioning (N=8,930)	Proposed Framwork (N=57)
a woman sitting on a couch	the talbe has a white	Ki Jung and Ki Taek sitting at a table
the floor is made of wood	the pizza on table	Ki Taek in a blue shirt and Ki Taek holding a glass of wine
a glass of water on the table	the table in room	Ki Jung sitting on a couch and Ki Jung holding a glass and Ki Jung with brown hair
a hand hoding a knife	the woman has a hair	Chung Sook sitting on a couch
green and white bad	the table in front of a white wall	the shirt is green
man in a blue shirt	the table has a black hair	a wooden table and a black and silver table
a woman hoding a glass of wine	the hand has a black hair	a chair in the background and a gray chair
the table is wooden	the glass has a white	a bottle of wine and a black bottle of wine
a chair in the background	the hand on a man	a man holding a glass of wine
a chair in the room	the light on wall	a plate of food and a white plate on a table
...	...	...

Figure 13. Comparison of captions between the existing dense captioning model and the proposed framework.

post-processing to convert the captions into a form suitable for understanding the movie. In addition, if the number of region captions to be extracted is limited, a smaller number of region captions can be generated.

As shown in the above study, the existing DenseCap models, which lack a concept of the importance of area captions for each scene, generate large numbers of captions 105 on average (the number of area captions proposed by the DenseCap researchers).[20] Therefore, from the point of view of summarizing the information in each movie scene, the region captions generated and selected by the algorithm proposed in this study are much more efficient. In other words, it is a more efficient method for the summary and management of movie information through image captioning when only a minimal amount of captioning information can be stored and managed according to the order of importance.

However, the core is in whether the top five captions selected by the proposed algorithm properly extract the important regions for each image, as intended by the director. In other words, it is necessary to additionally verify the appropriateness of the critical region captions selected by the proposed algorithm.

For such verification, it is necessary to qualitatively confirm the intention and opinion of the film production director, but there are practical limitations to this, such as having to make contact with the director. Besides that, since this verification method is difficult in terms of the usability of the proposed algorithm in the future, a suboptimal solution is to investigate what customers who have watched "Parasite" judge to be important region captions on each screen. The importance judged by the proposed algorithm can be evaluated as reasonable if the captions that many customers judge to be important are the same as those extracted by the proposed algorithm. To evaluate the validity of the region caption results by the proposed algorithm, a qualitative evaluation was performed on the images of 50 major scenes from the movie targeting 30 people who had watched the movie. An average of 105 region captions were obtained from each scene of 7922 movie images, and the average number of region captions corresponding to the top 10% based on importance was 10.5. Therefore, this criterion was applied to all 50 scenes, and the region caption corresponding to the top 10% was selected based on the importance of each scene. Then, 30 survey respondents were asked to evaluate whether these selected region captions were judged to be important in each scene. First, the survey respondents evaluate whether each of the top 10 captions selected as important for each of the 50 scenes is judged to be important (important=1/not important=0) in that scene. Then, the ratio of the total frequency judged to be important divided by the total number of captions for each scene was obtained, and then the average value of 30 respondents was obtained. In other words, the meaning of 90.0% in scene 1 is the value that each respondent evaluated for the caption provided in the scene divided by the total number of respondents. "A region caption was evaluated as being important if more than half of the respondents judged that caption as being important," and the accuracy was calculated as the ratio of the number of important region captions to the total number of region captions extracted for each scene. Table 3 lists the accuracy of each scene according to this qualitative

Table 3. Qualitative evaluation results of 30 people.

Scene	Accuracy (%)	Scene	Accuracy (%)	Scene	Accuracy (%)
1	90.0	18	96.5	35	53.8
2	88.9	19	85.7	36	66.7
3	90.9	20	97.3	37	70.0
4	78.6	21	98.0	38	92.9
5	84.6	22	75.0	39	82.8
6	97.4	23	71.4	40	81.6
7	58.3	24	64.3	41	75.0
8	98.7	25	63.6	42	92.3
9	76.9	26	83.3	43	87.5
10	80.8	27	97.4	44	92.4
11	55.4	28	91.7	45	98.1
12	87.6	29	81.5	46	90.9
13	91.0	30	90.1	47	78.6
14	91.3	31	71.2	48	81.8
15	87.5	32	50.0	49	77.5
16	87.5	33	95.2	50	97.6
17	81.8	34	63.6	—	—

survey. The average accuracy was 80.8%, thus indicating that the proposed algorithm identified important region captions at a very meaningful level.

Conclusion

This study proposes an image captioning framework centered on a core region for the automatic understanding of movies, which is one of the story-driven video contents. In the proposed framework, the character identification module was designed and implemented to reflect character information, which is an important element of a movie, in the area caption importance calculation and post-processing process. The HOG algorithm, which has been shown to have an accuracy of 98.8%, was adopted as the algorithm for generating the character face image data set. Meanwhile, the EfficientNet-B5 model, which has been shown to have 99.85% facial recognition accuracy, was selected for use after comparing different CNN models for character face recognition. The proposed framework calculated the importance of the region caption while considering four variables affecting the region caption importance: "region box confidence score," "region area," "distance between region and scene center," and "object type." It was confirmed that the algorithm proposed in this study reduced unnecessary region captions that appeared in the existing dense captioning model and successfully identified region captions that are relatively important for understanding movie scenes. In addition, the number of region captions was further reduced by integrating region captions describing the same object, and words referring to people in the region caption were converted into character names so that the region captions could reflect character information. As a result, it was confirmed that the result of the proposed framework extracted a caption that was more meaningful for understanding the movie scene than the caption of the existing dense captioning module, and it contained many cinematic elements. This study is meaningful in that it analyzed story-driven video contents, which comprise of a high-level domain that has not been dealt with in the existing image captioning domain. The findings of this research can also be used in various fields such as annotation and abstraction, retrieval, and recommendation of movies. In addition, if only the caption of the core region is considered in an application field where real-time analysis is needed, it is expected to contribute to reducing the computation and processing time.

However, the framework proposed in this study has a limitation in that it does not consider the relationship between objects and scenes by analyzing individual region captions and single scenes. To solve this limitation, we intend to conduct research on building scene graphs and ontology by identifying relationships between scenes and defining objects, object properties, and object relationships from region captions. Further studies will also be conducted to ensure that the proposed framework shows consistent performance in films in various genres.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Hyesun Suh https://orcid.org/0000-0001-9469-7707

References

1. Farhadi A, Hejrati M, Sadeghi MA, et al. Every picture tells a story: generating sentences from image. In: Computer Vision-ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010, 15–29.
2. Kuznetsova P, Ordonez V, Berg TL, et al. TREETALK: composition and compression of trees for image descriptions. *Trans Assoc Comput Linguistics* 2014; 210: 351–362.
3. Mao J, Xu W, Wang J, et al. Explain images with multimodal recurrent neural networks. arXiv preprint 2014, 1410.1090.
4. Polina K, Vicente O, Berg AC, et al. Collective generation of natural image descriptions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Volume 1. Association for Computational Linguistics, 2012, pp. 359–368.
5. Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguistics* 2014; 2: 67–78.
6. Oriol V, Alexander T, Samy B, et al. Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015. 3156–3164.
7. Xu J, Efstratios G, Basura F, et al. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015. 2407–2415.
8. Junhua M, Jonathan H, Alexander T, et al. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27-30 June 2016, 11–20.
9. Hao F, Saurabh G, Forrest I, et al. From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015, 1473–1482.

10. Shubo M and Yahong H. Describing images by feeding LSTM with structural words. In Multimedia and Expo (ICME), 2016 IEEE International Conference on. IEEE, Seattle, WA, USA, 11-15 July 2016, 1–6.
11. Minsi W, Li S, Xiaokang Y, et al. A parallel-fusion RNN-LSTM architecture for image caption generation. In: Image Processing (ICIP), 2016 IEEE International Conference, Phoenix, AZ, USA, 25–28 September 2016, 4448–4452.
12. Girish K, Visruth P, Sagnik D, et al. Baby talk: understanding and generating image description. In Proceedings of the 24th CVPR, Colorado Springs, CO, USA, 20–25 June 2011.
13. Kim D. J, Choi J, Oh T. H, et al. Dense relational captioning: triple-stream networks for relationship-based captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019, 6271–6280.
14. Peyre J, Sivic J, Laptev I, et al. Weakly-supervised learning of visual relations. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017, 5179–5188.
15. Plummer B. A, Wang L, Cervantes C. M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence model. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 Dec. 2015, 2641–2649.
16. Yang L, Tang K, Yang J, et al. Dense captioning with joint inference and visual context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017, 2193–2202.
17. Yikang LI, Wanli O, Bolie Z, et al. An efficient subgraph-based framework for scene graph generation. *Computer Vision –ECCV* 2018: 346–363.
18. Zakir H, Ferdous S, Mohd FS, et al. A comprehensive survey of deep learning for image captioning. *ACM Comput Surv* 2019; 51: 1–36.
19. Krishna R, Zhu Y, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int Journal Computer Vision* 2017; 123(1): 32–73.
20. Johnson J, Karpathy A and Fei-Fei L. DenseCap: fully convolutional localization networks for dense captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016, 4565–4574.
21. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: 2005IEEE Computer Society Conference Computer Vision Pattern Recognition (Cvpr'05), San Diego, CA, USA, 20–25 June 2005; 1: 886–893.
22. Tan M and Le QV. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint 2019, 190511946.
23. Justin J, Andrej K and Li FF. DenseCap: fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016, 4565–4574.
24. Linjie Y, Kevin T, Jianchao Y, et al. Dense captioning with joint inference and visual context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017, CVPR) 2016;1978–1987.
25. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint 2014; 1409.1556.