Journal of Advanced Marine Engineering and Technology, Vol. 45, No. 6, pp. 434~441, 2021
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2021.45.6.434

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Description Region Expansion-based relationship-oriented dense image captioning model

Yeong-Jae Shin[1] · Seong-Beom Jeong[2] · Ju-Hyeon Seong[3] · Dong-Hoan Seo[†]

(Received December 1, 2021 ; Revised December 10, 2021 ; Accepted December 23, 2021)

Abstract: In dense image captioning, multiple sentences are generated by identifying the various contexts present in an image through the use of an object detection network to extract the regions of interest (ROIs) at which objects exist in the image. However, because the size of ROIs obtained from the object detection network is dependent on the size of the target object, each ROI is limited to containing only fragmentary information about the target object. Therefore, in this paper, we propose Description Region Expansion (DRE), in which clusters of ROIs are extracted from the detection network and combined into single ROIs. DRE is located behind the detection network. The expansion of the ROIs allows the relationships between object clusters rather than single objects to be expressed in dense image captioning. To verify the validity of the model, training and evaluation were conducted using the Visual Genome dataset. The BELU-1 score of 0.678 and Meteor score of 0.463 achieved by the model confirms its excellent performance.

Keywords: Dense image captioning, Object detection network, Description region expansion, Region of interest

1. Introduction

The increase in CCTV density and technological advances in video equipment have made machine-based surveillance systems possible, and, more recently, common. The installation of CCTVs in almost all public areas has revealed limitations in manned monitoring systems resulting from the vast number of CCTVs, including invasion of privacy. Various studies on unmanned surveillance systems to solve these problems are being actively conducted. Recently, the development of object recognition technology has enabled various applications in image understanding through the utilization of information contained in images. Image captioning refers to the technology of generating sentences that describe the situations in images. It is an essential element of unmanned surveillance systems, but research on this topic is limited.

Image captioning is a technology for describing object information and context in an image. Because this technology includes the processes of both image analysis and processing as well as the generation of natural language descriptions, it involves the fusion of two disparate technologies. Methods for generating one description for each image have been intensively studied in image captioning. However, this approach is not appropriate because applications such as Question Answering **[1]**, which includes unmanned surveillance systems, requires all the objects in the image to be described. However, the focus of mainstream research is still on general image captioning and there is a lack of interest in dense image captioning **[2]**, which requires the acquisition of various information.

Methods for selecting image descriptions from a set of predefined sentences or corresponding words in a fixed template have previously been used for image captioning **[3]-[5]**. However, these methods do not provide smooth descriptions of the given image because the generated sentences are limited. Accurate sentence generation has been made possible by the introduction of deep learning technology, which has led to significant achievements in computer vision and natural language

† Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics & Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412

1 M. S., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: yjshin0329@gmail.com, Tel: 051-410-4822

2 M. S., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: sincere96@g.kmou.ac.kr, Tel: 051-410-4822

3 Assistant Professor, Department of Liberal Education & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: jhseong@kmou.ac.kr, Tel: 051-410-5031

processing. In particular, the encoder-decoder-based model achieves excellent performance by combining a convolutional neural network (CNN) for inferring the context of the image with a recurrent neural network (RNN) for generating the sentence description based on the feature vector. The use of this structure has become mainstream because image captioning models can be easily designed and debugged owing to the intuitive modular structure modularized with encoders and decoders.

Vinyals *et al.* **[6]** were the first to apply the encoder-decoder structure to image captioning. They extracted the feature vector of the image using Inception v3 as the encoder, and generated sentences using a long-short term memory (LSTM)-based decoder.

Xu *et al.* **[7]** and You *et al.* **[8]** proposed an attention technique that focused on visually important parts of images such as humans. The attention technique achieved excellent performance by using a structure that allowed for visual and word-level attention. He *et al.* **[11]** proposed a model to improve structural accuracy by focusing on the elements of the sentence to estimate the part-of-speech during the sentence generation process. Although the generation of image caption sentences that could accurately describe the context was demonstrated in these studies, only one caption was generated for each image. Unlike a caption dataset, a typical image often contains multiple contexts in one scene. Therefore, studies on dense image captioning for the simultaneous description of numerous contexts, such as those in the actual environment, are being conducted.

Johnson *et al.* **[2]** proposed Fully Convolution Localization Networks (FCLN) to generate multiple sentences from one image based on Faster-RCNN. The developed model could output information on all the objects by using a dedicated dataset for dense image captioning and combining it with the latest object detection algorithm. Yang *et al.* **[9]** and Li *et al.* **[10]** improved the network structure to use only the feature vectors. They also improved the accuracy of the image masks for captioning by including context features. Although these models could efficiently acquire detailed information by focusing on individual objects, they had weak capabilities for describing the relationships between objects required by humans.

To solve this problem, we propose a relation-oriented dense image captioning model in this paper. The proposed model is based on Description Region Expansion (DRE) and generates captions for each object cluster. Objects are extracted using Yolo v3-based region proposals for fast object detection. The limitations of existing captioning techniques based on the feature vector of a single object are overcome by estimating a cluster object given by the intersection over union (IOU) of the objects in an image and their adjacent objects. The proposed model generates captions for each cluster object by applying an injection-based image captioning network based on the feature vector of the estimated bounding box. Through this process, relationship-oriented sentence generation is made possible because the feature vector includes a bounding box that covers the surrounding objects in place of simply acquiring surrounding information through a margin or preprocessed context information. To verify the validity of the proposed model, training and testing were conducted using the Visual Genome dataset **[19]**.

2. Related Studies

2.1 Encoders in Image Captioning

The image captioning model is divided into two steps, namely, image feature vector extraction and sentence generation from the feature vectors. In image captioning, information is generally acquired immediately before the classifier in the form of the image feature vector of the image without using class information from object recognition. Therefore, despite the application of image analysis models, the use of object information from object recognition models in most image captioning models varies with the characteristics of the image captioning model. In such image captioning models with an encoder-decoder structure, the expressive power of the sentence is dependent on the performance of the encoder. Therefore, a precise image captioning model is required to classify the various objects and an accurate object recognition model should be used for the encoder.

Early studies on deep learning have focused on object recognition and resulted in the development of several excellent object recognition models. In particular, the VGGNet **[12]**, ResNet **[13]**, Inception v3, and Inception v4 **[14][15]** models achieved excellent performance in the ImageNet Large Scale Visual Recognition Competition (IRSVRC), and have been applied as encoders for image captioning. The problem of gradient loss due to the deepening of the CNN layer was solved through the use of the residual block in ResNet, which has been widely applied in image captioning models. The application of object detection models such as RCNN and Yolo, which have relatively complex structures, was limited compared to existing object recognition models because of their low object classification and computational efficiency. Image captioning models that use object

detection models capable of detecting even small and diverse objects are being studied. An object detection model such as Yolo **[16][17]** or Faster R-CNN **[18]** is typically used as an encoder. Faster-RCNN has been widely used as an encoder because it can offer more precise detection than Yolo and has a relatively easily modified network. However, Faster-RCNN has a disadvantage in that its detection speed is relatively slow compared to Yolo.

2.2 Dense Image Captioning

Unlike conventional image captioning, multiple captions are acquired from a single image in dense image captioning. Therefore, it is necessary to provide information on the various objects to the decoder in the encoder stage. However, because only a single feature vector over the entire image is input to the decoder in existing methods, the trained model can only generate limited captions. To solve this problem, approaches for generating and inputting feature vectors through the separation of object units in the image have been studied.

Johnson *et al.* **[2]** applied an object detection model that output the location information of the various objects that existed in the image together so that their feature vectors could be input. In their FCLN model, a feature vector for each object region was extracted by estimating the relative position in the image feature vector based on the positions of the bounding boxes generated by Faster-RCNN. The caption for each object was generated by performing individual captioning on each extract feature vector.

Subsequently, Yang *et al.* **[9]** and Li *et al.* **[10]** separately analyzed the contextual features in the entire image to obtain the background and context information lost during feature vector extraction. Although this approach could capture the interaction of an object with the surrounding space, it still had the limitation of outputting only local information. This is because only extremely local information was provided to the decoder when only a single object was acquired in the process of extending the feature vector. Therefore, to solve this problem, information on complex object clusters should be provided.

2.3 Evaluation Metrics for Image Captioning

The performance of image captioning technology is commonly evaluated using performance indicators for natural language processing machine translation. The most commonly used performance indicators are the Bi-Lingual Evaluation (BLEU) **[20]** and Metric for Evaluation of Translation with Explicit Ordering (METEOR) **[21]**.

2.3.1 Bi-Lingual Evaluation Understudy (BLEU)

The BLEU score provides a quantitative indication of the degree of agreement between the correct answer sentence and the generated sentence based on the n-gram, which is a sequence of n consecutive words. The BLEU score can be calculated as BLEU-1, which is based on a single word, up to BLEU-4 using four words depending on the number of words to be compared. The BLEU score is calculated as follows:

$$\text{BLEU} = \text{Brevity} \times \exp\left(\frac{1}{N}\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

$$\text{Brevity Penalty} = \begin{cases} 1 & if \;\; c > f \\ \exp\left(1 - \frac{r}{c}\right) & if \;\; c \leq f \end{cases} \quad (2)$$

As shown in **Equation (1)**, the BLEU score is calculated using the Brevity Penalty (BP), $\log p_n$, which is the accuracy for each n-gram; w_n, which is the weight of the n-gram; and N, which is the number of n-grams. The total sum of w_n should be 1. Therefore, w_n is 1 for 1-grams and 0.25 for 4-grams. The BP prevents the short length of short generated sentences from increasing the accuracy score spuriously. If the length c of the sentence generated by the model is less than or equal to the length r of the actual sentence, a penalty is imposed as given in **Equation (2)**. A high BLEU-1 score indicates that the generated sentence uses the same words as the correct sentence, while a high BLEU-4 score indicates that the generated sentence uses similar expressions to the correct sentence.

2.3.2 Metric for Evaluation of Translation with Explicit Ordering (METEOR)

In general, people may often use different words with the same meaning to describe scenarios. For example, the BLEU score is low for the two sentences “The man is holding a cellphone” and “A boy holding the smartphone” even though they have similar meanings. This is because synonymous synonyms are not considered when calculating the BLEU score. Therefore, the Meteor score was calculated by considering the morphemes and synonyms.

The Meteor score is based on the harmonic average of the precision and recall of the 1-gram. It is calculated using the ratio of the matching words between non-n-gram generated sentences and the correct sentences. A penalty is imposed for long sentences:

$$\text{Meteor} = F_{mean}(1 - penalty) \quad (3)$$

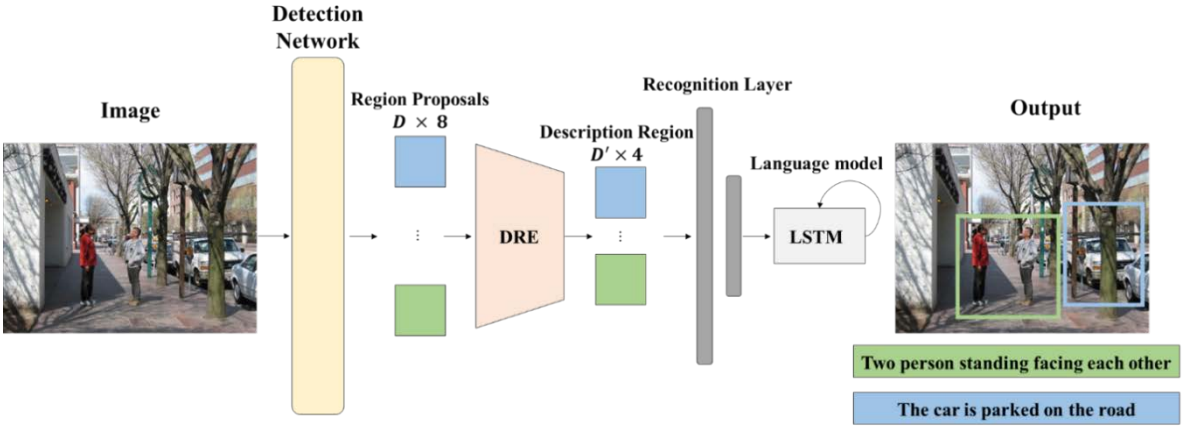

Figure 1: The structure of the proposed image captioning model

where the imposed penalty depends on the sentence length, and F_{mean} is the harmonic average of the 1-gram precision and recall considered for each word.

3. Proposed Dense Image Captioning

3.1 Overview of the Proposed Method

Although existing dense image captioning methods can generate captions for all the objects in an image and present detailed information such as the object properties, they can only output very local information. To solve this problem, we propose a DRE-based, relational-oriented dense image captioning model in which clustered objects that integrate high-relevance surrounding objects are extracted. Generally, a sentence is output in dense image captioning through a decoder by cropping the feature vector of the image based on the bounding box of the object generated by the object detection model. In this process, only the minimized object area can be obtained owing to bounding box regression in the object detection model. The surrounding background and related objects are removed during feature vector extraction based on the minimized object area. To solve this problem, the description area is expanded to the union between the detected object areas in the proposed model, and a feature vector that includes the surrounding objects is output. **Figure 1** shows the overall system architecture of the proposed model.

The detection network uses a pre-trained Yolo v3 **[15]** object detection algorithm for region of interest (ROI) prediction. The input image is first converted to $3 \times 416 \times 416$ and input to Darknet53, the backbone of Yolo v3, for feature vector extraction. Darknet53 consists of 52 convolution layers and one average pooling layer. The size of the feature map is adjusted by setting the stride of the convolution layer without using a max-pooling layer. The $3 \times 416 \times 416$ input image passes through DarkNet and is converted into a $1024 \times 13 \times 13$ feature vector. (x_1, y_1, x_2, y_2) bounding boxes are finally output by the subsequent detection layer. Each ROI contains eight predicted quantities, namely, the image index, bounding box information (x_1, y_1, x_2, y_2), objectness score, confidence score, and classified class. The bounding box information contains a total of four values (x_1, y_1, x_2, y_2) where (x_1, y_1) are the coordinates of the upper left corner of the predicted ROI, and (x_2, y_2) are the coordinates of the lower right corner. The objectness score is the probability that an object exists in the predicted ROI, and the confidence score quantifies the classification probability. The proposed model integrates the surrounding objects using only the coordinates and objectness of the ROIs.

3.2 Detection Network

Because the size of the ROI obtained from the detection network is tailored to the size of the target object, it is difficult to obtain the relationship between different objects with only a single ROI. It is therefore necessary to expand the ROI by combining adjacent bounding boxes. To achieve this, we propose DRE, which expands the ROI by identifying objects with relationships to one another based on their IOUs.

To describe an object in a sentence, the area of interest must include the object and its surrounding environment. Therefore, a margin of 10 % is added to each bounding box obtained by preprocessing to expand the size of the box. Because some bounding boxes obtained during this process may be too small to be expressed as sentences, bounding boxes below a certain size are deleted. **Figure 2** illustrates the DRE process. The remaining bounding boxes are first sorted in descending order of their objectness score. The DRE process proceeds in the following

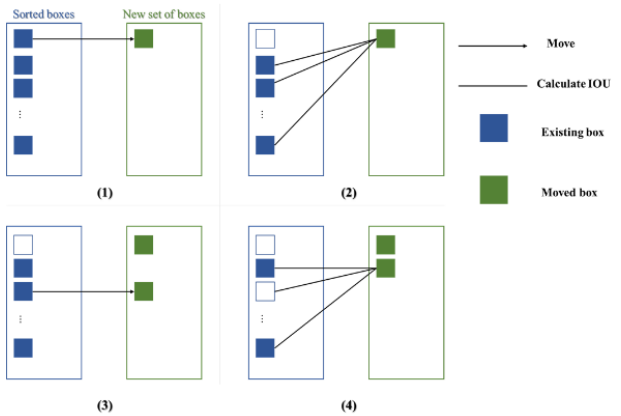

Figure 2: The sequence of Description Region Expansion

sequence, where each step is graphically represented in **Figure 2**:

1. After creating a new set of bounding boxes, the bounding box with the highest score is moved.
2. The IOU between the moved and remaining bounding boxes is calculated to understand the relationship between objects.
3. Objects with an IOU of 0.05 or higher calculated in the previous process are recognized as related objects and moved to the set of bounding boxes.
4. Steps 2 – 4 are repeated until there are no more boxes to be moved.

After the above process is completed, the size of the new bounding box for each set is defined using minimum value of the upper left corner and the maximum value of the lower right corner of the set. That is, the entire set is converted into a bounding box. Therefore, D bounding boxes are reduced to D' bounding boxes. After the completion of the above process, each set becomes a large bounding box through the combination of bounding boxes included in the newly created set.

3.3 Bilinear Interpolation

The size and ratio of the bounding box output from DRE is adjusted to $3 \times 416 \times 416$ to fit the input size of the detection network. Therefore, it is necessary to adjust the coordinates of the bounding box to match the ratio of the image feature map bounding box.

The image feature map uses the output from the DarkNet53 backbone of the detection network. While the size of the feature map is $1024 \times 13 \times 13$, the extracted bounding box has the same ratio as the image, which has a size of 416×416. A 13×13 grid is therefore created, and the coordinates of the bounding box and mapped part cropped and extracted. The size of the feature map is then transformed into $1024 \times 7 \times 7$ using bilinear interpolation.

3.4 Recognition Network

The recognition network is composed of fully connected layers. It receives a $1024 \times 7 \times 7$ feature map as input. The input feature map is flattened into a one-dimensional vector and passes through a fully connected layer, which has a dimension of 512, and then through batch normalization before it is input to the language model.

3.5 Language Model

The language model receives the feature vector and embedded words as the input from the recognition network. The structure of the input x_t and output y_t of the language model are as follows:

$$x_{t-1} = feature\ vector\ , \tag{4}$$

$$x_t = S_t, t \in \{0 \dots N-1\}, \tag{5}$$

$$y_{t+1} = LSTM(x_t), t \in \{0 \dots N-1\} \tag{6}$$

The feature vector in **Equation (4)** is the output of the recognition network and is input only once at time $t = -1$. S_t in **Equation (5)** denotes an embedded word, and N the number of embedded words in the sentence. S_0 is the <start> token indicating the beginning of the sentence, and S_N is the <end> token indicating the end of the sentence. In **Equation (6)**, the input x_t is repeatedly input to the LSTM at each time point t to generate the output vector y_t and hidden vector h_t as $h_t, y_t = f(h_{t-1}, x_t)$ where f represents the LSTM. The size of each embedding vector and the hidden layer is 512. The loss function $L(S)$ of the language model is expressed as

$$L(S) = -\sum_{t-1}^{N} \log\left(p_t(S_t)\right) \tag{7}$$

where S_t is the input at each time point, and p_t is the probability distribution for all the words. $L(S)$ is calculated as the sum of the negative log likelihood at each time point. The parameters of the LSTM should be set to minimize $L(S)$.

4. Experiment and Results

4.1 Datasets and Preprocessing

The Visual Genome (VG) dataset was used in this study to train and verify the proposed model. The version of the VG dataset used contains a total of 108,000 images and 4.1 million sentences. Each sentence has a relation property that indicates whether the sentence describes one object or a relationship between objects. Here, to investigate the expression of relationships between the target objects in a sentence, only the cases where relational information exists were extracted from the 4.1 million sentences in the dataset. A total of 997937 sentences were used for training after the removal of short sentences consisting of four or fewer words. To optimize the dataset, we converted words with a frequency of five or less into tokens during the preprocessing process. In addition, the beginning and end of each sentence can be learned by adding a <start> token at the beginning of the sentence and an <end> token at the end. The total number of words, including the <unk>, <pad>, <start>, and <end> tokens, was 6655, which was used as the dimension of the decoder embedding and final output layers. The output dimension of the embedding layer for vectorizing sentence information was set to 512. This is the same as the dimension of the last fully connected layer constituting the LSTM and recognition layers.

4.2 Results and Discussion

The ROI is expanded to realize dense image capturing for describing the relationship between objects. The BLEU-1 to BLEU-4 scores, which compare successive identities with sentences in the dataset, were used to evaluate the performance of the proposed model. **Table 1** shows the BLEU-1 to BLEU-4 scores of the model. The proposed model achieved a lower BLEU-4 score than the compared models, which seems to imply that it could not generate accurate and rich sentences by focusing only on image information. However, it achieved the highest Meteor score of 0.463 while CAG-NET and FCNL achieved similar scores.

Table 1: The BLEU-1 to BLEU-4 and METEOR scores achieved by various models on the dataset

Model	BLEU-1	BLEU-2	BLUE-3	BLEU-4	Meteor
FCLN	-	-	-	-	0.305
CAG-NET **[23]**	-	-	-	-	0.316
ASG **[22]**	-	-	-	**0.176**	0.221
Proposed model	0.678	0.447	0.252	0.141	**0.463**

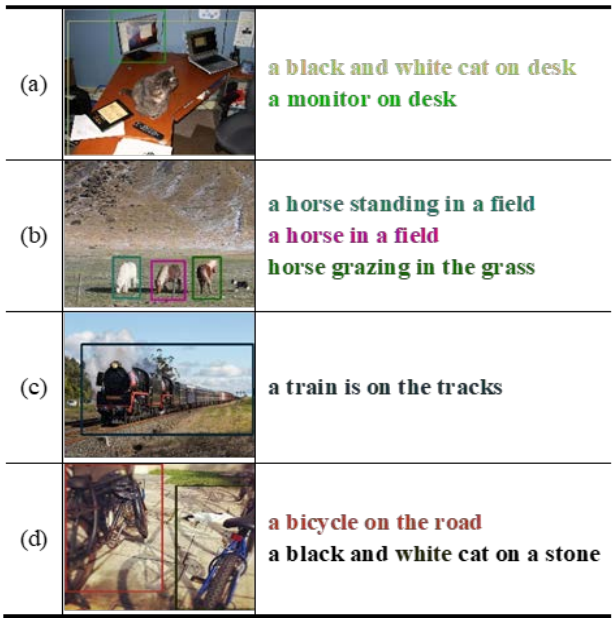

Figure 3: Samples of captions generated by the model

Figure 3 shows the sentences and their domains generated by the proposed model. **Figure 3(a)** shows an image of a cat sitting on a desk containing other objects, such as monitors and books. However, information on the objects other than the cat does not appear in the sentence because of the disappearance of the image feature vectors.

Figure 3(b) shows a scene of several horses in a field. A total of three description regions were generated, and the sentences describing the horses standing on the field or grazing on grass are well expressed. However, the detection network did not detect the object that looks like a dog on the far right because of its small size. A sentence for this object was therefore not generated.

Figure 3(c) shows a scene of a train on a track. No other objects except for the train are visible in the picture, and only one description region was generated. This sentence also accurately describes the train on the track.

Figure 3(d) shows an image of a parked bicycle and a cat. A total of two description regions were detected. The sentence for the left region accurately expresses information about the bicycle. For the right region, it was confirmed that the output sentence describes the cat in the back and not the bicycle.

5. Conclusion

In this paper, we proposed a relation-oriented dense image capturing model that creates sentences by concentrating on the relationship between the detected objects and expanding the range of sentence generation in dense captioning. In addition,

DRE was developed to create a description region that includes information on various objects by combining the ROIs extracted from the given image. Unlike existing dense captioning, the description region combined through DRE contains information about several objects. This makes it possible to express not only simple descriptions but also complex situations such as relationships and actions accurately and richly between objects. The performance of the proposed model was compared and evaluated using the BLEU score. A high performance score of 0.678 was achieved for the BLEU-1 score.

Through this study, it is expected that the relationships between objects will be more accurately described with richly expressive sentences in the field of dense captioning in the future.

Acknowledgement

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 21HCLP-C162922-01). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C1014024).

Author Contributions

Conceptualization, J. H. Seong; Methodology, Y. J. Shin; Software, Y. J. Shin and S. B. Jeong; Data curation S. B. Jeong; Writing-Original Draft Preparation, Y. J. Shin; Writing-Review & Editing, J. H. Seong and D. H. Seo; Supervision, D. H. Seo.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” In Proceedings of the 2015 IEEE international conference on computer vision, pp. 2425-2433, 2015.

[2] J. Johnson, A. Karpathy, and L. Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” In Proceedings of the 2016 IEEE conference on computer vision and pattern recognition, pp. 4565-4574, 2016.

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” In European conference on computer vision, Springer, Berlin, Heidelberg, pp. 15-29, 2010.

[4] Y. Yang, C. Teo, III, H, Daumé, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 444-454, 2011.

[5] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-grams,” In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220-228, 2011.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164, 2015.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, ... and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” In International conference on machine learning, Proceedings of Machine Learning Research (PMLR), pp. 2048-2057, 2015.

[8] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651-4659, 2016.

[9] L. Yang, K. Tang, J. Yang, and L. J. Li, “Dense captioning with joint inference and visual context,” In Proceedings of the 2017 IEEE conference on computer vision and pattern recognition, pp. 2193-2202, 2017.

[10] X. Li, S. Jiang, and J. Han, “Learning object context for dense captioning,” In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 1, pp. 8650-8657, 2019.

[11] X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, and W. Dong, “Image caption generation with part of speech guidance,” Pattern Recognition Letters, vol. 119, pp. 229-237, 2019.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556., 2014.

[13] 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, pp. 770-778, 2016.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” In Proceedings of the 2016 IEEE conference on

computer vision and pattern recognition, pp. 2818-2826, 2016.

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” In Thirty-first AAAI conference on artificial intelligence, 2017.

[16] J. Redmon, S. Divvala, R. Girshick and A. Farhadi “You only look once: Unified, real-time object detection,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016

[17] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.

[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, vol. 28, pp. 91-99, 2015.

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. Shamma, M. Bernstein and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” International journal of computer vision, arXiv preprint arXiv:1602.07332, 2016.

[20] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318, 2002.

[21] S. Banerjee, and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65-72, 2005.

[22] S. Chen, Q. Jin, P. Wang, and Q. Wu, “Say as you wish: Fine-grained control of image caption generation with abstract scene graph,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9962-9971, 2020.

[23] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, “Context and attribute grounded dense captioning,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6241-6250, 2019.