The Journal of Engineering

The 3rd Asian Conference on Artificial Intelligence Technology (ACAIT 2019)

IET Journals
The Institution of Engineering and Technology

EXplainable AI (XAI) approach to image captioning

eISSN 2051-3305
Received on 14th October 2019
Accepted on 19th November 2019
E-First on 27th July 2020
doi: 10.1049/joe.2019.1217
www.ietdl.org

Seung-Ho Han[1] ✉, *Min-Su Kwon*[1], *Ho-Jin Choi*[1]

[1]*School of Computing, KAIST, Daejeon, Republic of Korea*

✉ *E-mail: seunghohan@kaist.ac.kr*

Abstract: This article presents an eXplainable AI (XAI) approach to image captioning. Recently, deep learning techniques have been intensively used to this task with relatively good performance. Due to the 'black-box' paradigm of deep learning, however, existing approaches are unable to provide clues to explain the reasons why specific words have been selected when generating captions for given images, hence leading to generate absurd captions occasionally. To overcome this problem, this article proposes an explainable image captioning model, which provides a visual link between the region of an object (or a concept) in the given image and the particular word (or phrase) in the generated sentence. The model has been evaluated with two datasets, MSCOCO and Flickr30K, and both quantitative and qualitative results are presented to show the effectiveness of the proposed model.

1 Introduction

Image captioning, a subfield of computer vision (CV) and natural language processing (NLP), is the task of generating a textural description of a given image. Recently, deep learning techniques have been intensively used to this task with relatively good performance. Deep learning-based image captioning models normally use the encoder–decoder framework using convolutional neural network (CNN) and recurrent neural network (RNN). The encoder–decoder model consists of two phases: encoding and decoding. Normally a CNN-based encoder extracts the feature vector from the input image, then an RNN-based decoder generates a word for each time step. A sequence of words, i.e. a sentence, is generated as the caption [1, 2].

Due to the 'black-box' paradigm of deep learning, however, existing approaches are unable to provide clues to explain the reasons why specific words have been selected when generating captions for given images, as discussed in our earlier paper [3]. This limitation leads to generate absurd captions occasionally. To overcome this problem, Han and Choi [3] have proposed an explainable image captioning model, which provides a visual link between the region of an object (or a concept) in the given image and the particular word (or phrase) in the generated sentence.

The proposed model is shown in Fig. 1. Assuming that the model training has completed ('how to' will be presented in Section 3), the process of caption generation proceeds as follows. First, an input image is fed into the 'trained' model, which generates a caption and a weight matrix. Then, these caption and weight matrix are passed to the visualizer which highlights the major words appearing in the caption to their corresponding regions in the image. For the given image, the caption is generated using the language model trained with objects and words, and the weight matrix is produced by the attention model using the objects detected from the image and words in generated caption. The visualised final result shows several elements: colour-coded words in the generated caption, coloured region boxes on the image capturing the objects detected, and weight values for the word-region pairs of the same colour. Each weight value indicates the degree of relevance 'matching' between the word and the object in a word-region pair. These matched pairs provide the rationale why the caption was generated using the words selected.

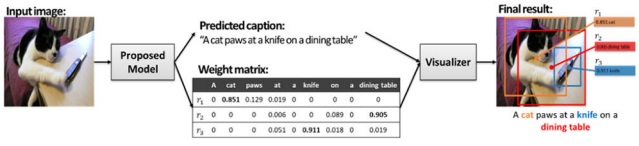

Fig. 1 *Process of image captioning and visualisation*

This paper is an extended version of the previous paper [3], that is, an eXplainable AI (XAI) approach to image captioning. The main contributions are as follows. First, we propose a novel image caption generator that can generate a more accurate caption by considering the region information and provide the visual explanation. Second, we propose a novel module for visual explanation, so-called 'explanation part', based on Bayesian inference. Third, through our experiments, we show quantitative and qualitative result of our model and verify the effectiveness of proposed model.

This paper is organised as follows. Section 2 provides related works. Section 3 presents the details of the proposed model. Section 4 presents the experimental results, and Section 5 concludes.

2 Related works

2.1 Image captioning with encoder–decoder model

Prior to using deep learning models, image captioning has been tackled by combining CV with NLP techniques. Deep learning techniques have improved the performance of image captioning, and especially deep recurrent models, called the 'encoder–decoder' models [4–6], have been adopted as the core of image captioning. In an encoder–decoder model, the encoder extracts a feature vector from an input image based on CNN, and the decoder generates a sentence using the feature vector based on RNN.

2.2 Image captioning with object detection

More recently, object detection algorithms have been used to obtain more detailed captions (or phrases) for specific parts of an image. Karpathy and Fei-Fei [7] proposed a deep visual-semantic alignment model that generated descriptions of images or region. This approach first calculates the scores for regions–words using an object detection algorithm [8], then trains the generative model using multi-modal RNN (m-RNN) [4] using image-caption data and pre-calculated scores. Using the trained model, a phrase is generated for an input region. Johnson *et al.*[9] proposed DenseCap to generate dense captioning (phrase descriptions) for selected regions, using fully convolutional localisation networks. The localisation layer proposes regions from an input image and

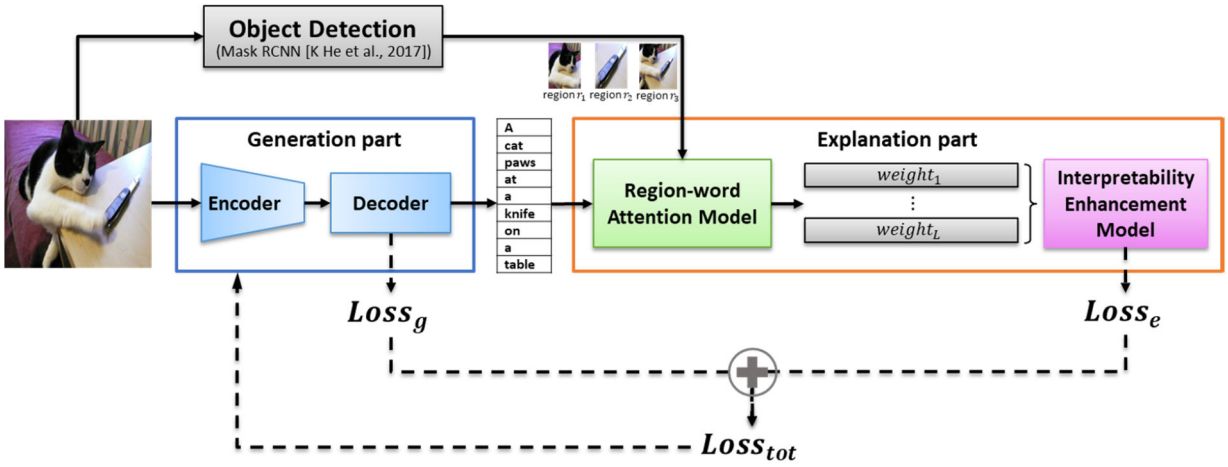

Fig. 2 *Architecture of proposed model*

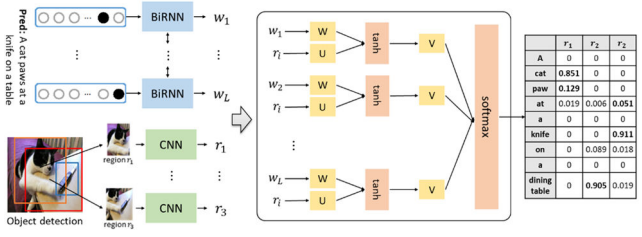

Fig. 3 *Region-word attention model*

extracts their features. Using these features, a RNN language model is trained, which generates short captions for selected regions as the final output.

2.3 Image captioning with attention mechanism

Neural processes involving attention have been chiefly studied in the computational neuroscience. In the last few years, many attention-based deep learning models have been studied in various fields, such as speech recognition, NLP, showing great performance. Recently, this concept of attention has been applied to image captioning task based on an encoder–decoder model. The encoder divides a given image regularly into grid regions and generates a set of feature vectors for the regions. Then these vectors are fed into an attention model, which assigns weights to the feature vectors. Finally, the decoder converts these feature vectors into context vectors by multiplying the weights from the attention model, then generates a caption using the context vectors. A good example of an image captioning model with attention layer is found in [10], which showed better performance than previous neural caption generators such as [1] which did not use an attention mechanism. This model [10] also highlights the very image part on which the attention layer focuses when generating each word. As other examples, Chen *et al.* [11] and Pedersoli *et al.* [12] proposed to use multiple attention models for spatial, activation, object etc., and showed better performance than single attention model.

3 Proposed model

This section describes the details of our proposed model. For the sake of completeness and readability, we repeat in Section 3.1, the same description from our earlier paper [3].

3.1 Model architecture

Fig. 2 shows our whole model architecture. In this figure, our model is divided into two parts: generation and explanation parts. The generation part generates the caption from given image using encoder–decoder architecture. The explanation part generates a weight matrix for regions in input image and words in generated caption. These parts also generate loss values, Loss_g and Loss_e. Both loss values affect the trainable parameters of generation part to consider region information. Details of each part are described as follows:

3.1.1 Generation part: The generation part is based on CNN-RNN encoder–decoder framework. The encoder extracts a feature vector for the full image, and the decoder generates the words using the feature vector. For the encoder, we use the VGG-16 [13] model and convert the size of all images into a fixed size to extract the image feature vector. For the decoder, we use the long-short-term memory (LSTM), which generates the words every time step using the image feature vector and word embedding. We also use a negative log likelihood loss function to jointly optimise the trainable parameters of the encoder–decoder model for image-caption pairs. However, this part cannot identify specific parts of the given image. Hence, we designed the explanation part in order for the generation part to consider the important objects that are detected from a given image when generating the caption and providing explanation from the generated caption.

3.1.2 Explanation part: The explanation part has two major roles depending on whether the generation part is in training or inferencing stage. During training, the explanation part generates Loss_e, an image-sentence relevance loss, which digitises whether the generated caption considers the objects in the input image well. The objects are extracted by using an object detection algorithm [14]. The more the generation part is trained, the better the model can generate a caption considering objects. During testing, the explanation part generates the weight matrix for the regions extracted from the input image and words generated from the generation part for the image. Each weight value represents the relevance between the object and the word in the pair. The highest weight values are taken in the final result as shown in Fig. 1. The explanation part has two components: (i) the region-word attention model and (ii) the interpretability enhancement (IE) model. The region-word attention model generates a weight matrix using the regions detected during object detection and the words in the generated caption. The IE model generates the image-sentence relevance loss using the weight matrix to assess whether a caption generated from the generation part well-reflects the objects.

3.2 Region-word attention model

Region-word attention model is a key component of the explanation part. Comparing with the visual attention model introduced in [8], our attention model has different training procedure to meet our purpose. The purpose of our attention model is to assist the generation part in considering region information. To achieve this, we use a concept of attention mechanism and our attention model generates a weight matrix for input regions and words. Fig. 3 shows our region-word attention model.

The left side of Fig. 3 represents the regions and words fed into the attention model. The regions are sub-images extracted from the original image by using the object detection algorithm [1]. The

words are generated from generation part during training stage. The middle of Fig. 3 represents the structure of the attention model. The attention model is parameterised as a feed-forward neural network, similar to other attention models. The right of Fig. 3 shows a weight matrix, which is an output of the model. In the weight matrix, each column represents weight vector for each region. Each weight value indicates the degree of relevance between each region and each word. The larger the value, the more relevant it is. Each weight value is computed as in (1).

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{L} \exp(e_{ik})}$$
$$e_{ij} = V \times \tan h\left(U \times r_i + W \times w_j\right) \quad (1)$$

where r_i is ith region ($1 \le i \le n$) and w_j is jth word ($2 \le j \le L$). The V, U, and W are trainable parameters to train the attention model. The weight, α_{ij}, represents a degree of relevance between w_j and r_i. The range of each α_{ij} is 0 to 1, and the sum of all of the α_{ij} is 1. The difference between our attention model and other attention models is a training procedure. Other attention models are jointly trained with an encoder–decoder model. However, our attention model is independently trained by using pre-trained embedding model based on vocabulary for caption dataset and all of the region labels. Details are described in the following section.

3.2.1 Training procedure of the region-word attention model: The explanation part is pre-trained before training the generation part. To pre-train the explanation part, the attention model is trained first, then an IE model is trained. In a training phrase, the inputs of the attention model are each region extracted from images and all words in ground-truth captions for the images. The output of the attention model is a weight vector for input region and words. To optimise the trainable parameters of the attention model, we use the mean-squared loss function by utilising the pre-trained embedding model with vocabulary in caption dataset and object categories for all the regions. Each weight value in a generated matrix is used as predicted value of mean-squared loss and word similarity between labels of region r_i and w_j is used as a truth value of mean-squared loss. Equation (2) shows a loss value for training the attention model.

$$\text{Loss}_{\text{att}} = \frac{1}{L}\sum_{k=1}^{L} (\text{similarity}(l_i, w_k) - \text{weight}(r_i, w_k))^2 \quad (2)$$

where L is the number of input words and l_i represents the label of r_i. The weight (r_i, w_k) is a weight value for region r_i and word w_k and the similarity (l_i, w_k) is a word similarity between the region label and word w_k. The similarity value is computed by using a pre-trained word-embedding, which constructed using dictionary and label categories. The range of similarity value is from −1 to 1. We use a value from 0 to 1 because −1 indicates that two words are semantically opposite in the embedding. As the model training progresses, therefore, the attention model is optimised to generate weight vector that each weight value is similar to embedding value.

3.3 Interpretability enhancement model

IE model generates the image-sentence relevance loss, Loss_e, using a weight matrix generated from the attention model. The IE model determines whether or not the generation part utilises the region information well when generating a caption. To this end, the model first picks the relevant region-word pairs that have highest weight value for each region in the weight matrix. The pairs are used as an input of the IE model. Using the region-word pairs, the IE model checks whether each region-word pair is actually correct what a region and word in the pair are related to actual data distribution. To do this, we use the concept of Bayesian inference [15]. The output of the IE model is a predicted posterior probability, $P(r_i|w_j)$ for a region given a word in the pair. This posterior probability means the region and word in selected pair are actually related in the actual data distribution. The reason for this confirmation is because the generated caption by generation part might be wrong during training of generation part. Therefore, if the posterior probability is high for the given region r_i and the word w_j, it means the word has high relevance to the region in the actual data distribution. In other words, the generated caption considered the r_i and w_j well. As a result, if the generated caption properly considers the all regions in a pair, the sum of posterior probabilities for the regions will be high, and the IE model will generate a low Loss_e. Otherwise, if the sum of posterior probabilities will be low, the model generates a high Loss_e and this loss value will affect the training of the generation part.

However, the posterior probability cannot be calculated directly, because we do not know the conditional distribution for the posterior probability. The target of conditional probability, w_j, is generated while generation part is being trained, whereas the IE model has to be trained before generation part. Hence, we use Bayesian inference to approximate the posterior probability using the prior probability and likelihood, based on the Bayes' theorem. Consequently, we can compute the image-sentence relevance loss (Loss_e) as shown in (3).

$$\text{Loss}_e = \sum_{i=1}^{n}\sum_{j=1}^{k}(1 - P(r_i|w_j))$$
$$P(r_i|w_j) = P(w_j|r_i) \times P(r_i) \quad (3)$$

where n is the number of regions in picked pair, and k is the number of selected words for each region. In our experiments, we use 1 or 2 for k. Loss_e is the image-sentence relevance loss. As shown in (3), the posterior probability is used to calculate Loss_e. As previously stated, to obtain the posterior probability, we use Bayesian inference with the likelihood, $P(w_j|r_i)$, and the prior probability, $P(r_i)$. As the likelihood and prior probability can be calculated statistically, these distributions are pre-calculated by using the training dataset. By approximating the posterior probability, we can compute the Loss_e by adding all values that each posterior probability subtracted from 1 for the all pairs. At the end, final loss value is passed on to the generation part so that the generation part considers the region information when generating a caption.

As region r_i and word w_j are in the form of vectors, we cannot directly obtain all probability values. Therefore, we design a model that is fed the region and word vectors, returning the posterior probability as shown in Fig. 4. This model is parameterised as a feed-forward neural network. The region-word pairs are selected from weight matrix and used in IE model. To train the IE model, we use the cross-entropy loss function. The truth value in loss function is the product of the likelihood and prior probability and predicted value is generated posterior probability from the IE model.

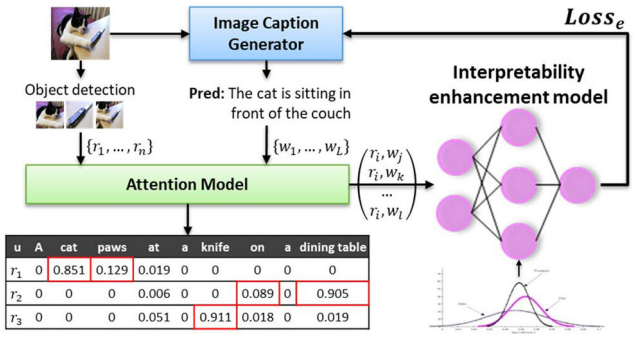

Fig. 4 *Interpretability enhancement model*

Table 1 Quantitative results for MSCOCO dataset, comparing our model with other models

Model	B@1	B@2	B@3	B@4	METEOR
[4]	67	49	35	25	—
[1]	66.6	46	32.9	24.6	—
[10]	71.8	50.4	37.5	25	23
[7]	62.5	45	32.1	23	19.5
[20]	—	—	37.2	27.6	24.7
[11]	71.9	54.8	41.1	31.1	25
ours (P1)	72.5	55.1	38.8	28.3	26.2
ours (P2)	71.1	53.1	39.4	29.5	24.7

High is good in all columns and the numbers in bold face are the best-known results and (—) indicates unknown scores.

Table 2 Quantitative results for Flickr30K dataset, comparing our model with other models

Model	B@1	B@2	B@3	B@4	METEOR
[4]	54.7	23.9	19.5	—	—
[1]	66.3	42.3	27.7	18.3	—
[10]	66.9	43.9	29.6	19.9	18.5
[7]	57.3	36.9	24	15.7	—
[20]	—	—	30.2	21	19.2
[11]	66.2	46.8	32.5	22.3	19.5
ours (P1)	67.4	47.5	31.7	21.4	19.9
ours (P2)	65.9	45.8	32.2	22.1	18.6

4 Experiments

4.1 Experimental setting

4.1.1 Dataset: For image captioning, we used two benchmark datasets: MSCOCO [16] and Flickr30K [17]. The MSCOCO (2014) contains 82,783 images in the training set, 40,504 images in the validation set, and 40,775 images in the test set. The Flickr30K consists of 31,783 images with 158,915 crowd-sourced captions, and we used it by splitting 29,000 images for training, each 1000 for validation and for testing, for fair comparisons to our baseline paper [7]. Each image in all datasets comes with five descriptive captions written by human.

4.1.2 Data pre-processing: For training our proposed model, we pre-processed the datasets to make the model operate as we intended and to maximise performance. In the case of the caption data, we converted all sentences to lower case, discarded non-alphanumeric characters, and removed all captions for >15 words. We also filtered the words to those that occurred too frequently, such as '*the*', '*this*', etc., and we used a fixed vocabulary size including the labels for all regions. In the case of the image data, before we pre-processed the image data, we constructed the region dataset. To construct the region dataset, we used several principles as follows. We used the region larger than confidence level of 85% or more from object detector. We discarded regions smaller than 50×50 pixels. Then, we pre-processed the whole image data. We altered the size of the all images to the same size (256×256). We also discarded images having no regions with a confidence level of 85% or more.

4.2 Quantitative analysis: image captioning

To evaluate the results of image captions generated from our proposed model, we use the evaluation metrics, BLEU [18] and METEOR [19] scores. The BLEU and METEOR scores are an algorithm for the evaluation of sentence generated by machine. BLEU@*n* (B@*n*) represents the geometric average of the n-gram precision. METEOR is based on the harmonic mean of unigram precision and recall, and it also considers several features such as stemming and synonymy-matching, with the standard exact word-matching. We evaluate the image caption results with these evaluation metrics by comparing ours to other caption-generation models such as m-RNN [4], NIC [1], NIC with visual attention [10], deep visual-semantic alignments [7], attention correctness [20], and SCA-CNN [11]. In the case of our model, we experimented with two cases in accordance with the number of selected word pairs for each region in the IE model (referred as model with P1 and P2).

4.2.1 Analyse the B@1 and B@2 scores in our model with P1: As shown in Tables 1 and 2, our model with P1 outperforms other models for all datasets for B@1 and B@2 scores. The reason for these results is that our model is trained to generate a caption reflecting the important regions in given image. The generated captions from our model tend to include at least one related word for each region; consequently, the captions contain words as many as the number of regions found. Considering the B@1 and B@2 scores, the performance of generated sentences in accordance with unigram and bigram precision was evaluated. Using BLEU metrics for image captioning tasks might penalise some correctly generated sentences. Thus, the captions that reflect salient objects as much as possible are advantageous for receiving high BLEU scores. Therefore, our model with P1 received higher B@1 and B@2 scores than others for all datasets.

4.2.2 Analyse the B@3 and B@4 scores in our model with P1: The scores of B@3 and B@4 for our model with P1 are second, except for P2, as in Tables 1 and 2. In the same context as the B@1 and B@2 cases, the scores of B@3 and B@4 were calculated by comparing more words. However, this caused the advantage of captions reflecting the important objects diminish. Thus, because the model with P1 was trained with a Loss_e considering only one relevant word, it implicitly ignores the relations between words related to each region. As a result, the scores of B@3 and B@4 decreased. Besides, the model with P1 cannot consider the semantic aspect when generating a caption; therefore, it cannot cover the relation between regions or words. This is a limitation of our model.

4.2.3 Compare our model with P1 and model with P2: For all datasets, the B@1 and B@2 scores of P1 are higher than P2. The reason for this is same as the reason explained above. In the case of the B@3 and B@4 scores, the model with P2 is higher than the model with P1. With P2, this is influenced by the Loss_e generated from the IE model, which was trained using two region-word pairs for each region. Thus, the model with P2 reflected the two words for each region when generating its caption, consequently, this has an advantage when evaluating with more words (B@3 and B@4 cases).

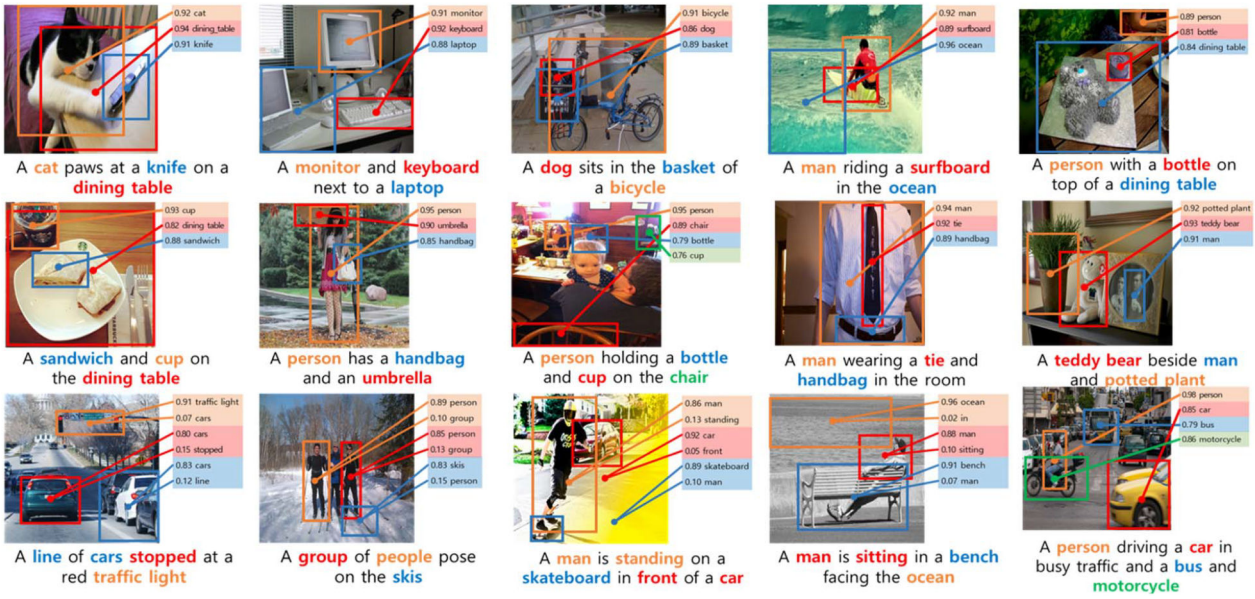

Fig. 5 *Examples of final results generated from proposed model*

4.3 Qualitative analysis: final result

We next show our final results that contain generated captions for the given images and the visual explanation by colouring and boxing regions with the same colour as related words in the caption, as shown in Fig. 5. For each result in this figure, there are image with coloured boxes that indicate the detected regions, generated caption that some words are coloured in same colour as regions and relation information (weight values) next to each image. The coloured words indicate that each word was influenced by the region with the same colour. The first and second rows in this figure show the results of our model with P1 from Tables 1 and 2 and third row represents the results of our model with P2. This means that each caption in each result considered only one word for each region in the pairs. In the first result, a generated caption is 'A cat paws at a knife on a dining table'. In this caption, the coloured words, 'cat', 'knife', and 'dining table', are generated by considering the regions with the same colour based on the weight values. For example, the word 'cat' is coloured orange, and the orange region box encircles the cat with a weight value of 0.92. Thus, because the generated caption reflected the region information, we can connect the regions and specific words. Besides, the model provides a visual explanation for why the words were selected.

5 Conclusion

In this paper, we proposed the explainable image caption generator, which generates a caption by considering the region information for a given image and provides explanation for why the words in the generated caption are selected. To this end, we designed an explanation part that composed of the region-word attention model and IE model. Using these models, this part generates an image-sentence relevance loss that influences the generation part during the training stage and generates a weight matrix representing the relations for the regions extracted from the given image and words in the generated caption. In our experiments, we analysed the quantitative results for generated captions by comparing our model to others. We also showed the qualitative results of proposed model. In the future, we plan to improve our model by solving our limitations. In particular, we will develop a semantic attention module that can discover attributes for a given image and utilise it together with the region-word attention model.

6 Acknowledgments

This research was supported by Korea Electric Power Corporation. (Grant number: R18XA05).

7 References

[1] Vinyals, O, Toshev, A, Bengio, S, *et al.*: 'Show and tell: a neural image caption generator'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 3156–3164

[2] Kiros, R, Salakhutdinov, R, Zemel, R.S..: 'Unifying visual-semantic embeddings with multimodal neural language models', arXiv preprint arXiv, November 2015, 1411.2539

[3] Han, S, Choi, H.: 'Explainable image caption generator using attention and Bayesian inference'. Proc. Int. Conf. Computational Science and Computational Intelligence (CSCI), 2018

[4] Mao, J, Xu, W, Yang, Y, *et al.*: 'Explain images with multimodal recurrent neural networks', arXiv preprint arXiv, October 2014, 1410.1090

[5] Donahue, J., Anne Hendricks, L., Guadarrama, S., *et al.*: 'Long-term recurrent convolutional networks for visual recognition and description'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 2625–2634

[6] Chen, X., Lawrence Zitnick, C.: 'Mind's eye: a recurrent visual representation for image caption generation'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2422–2431

[7] Karpathy, A, Fei-Fei, L.: 'Deep visual-semantic alignments for generating image descriptions'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 3128–3137

[8] Girshick, R, Donahue, J, Darrell, T, *et al.*: 'Rich feature hierarchies for accurate object detection and semantic segmentation'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014, pp. 580–587

[9] Johnson, J, Karpathy, A, Fei-Fei, L.: 'Fully convolutional localization networks for dense captioning'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016, pp. 4565–4574

[10] Xu, K, Ba, J, Kiros, R, *et al.*: 'Show, attend and tell: neural image caption generation with visual attention'. Proc. IEEE Int. Conf. Machine Learning (ICML), Jun, 2015, pp. 2048–2057

[11] Chen, L, Zhang, H, Xiao, J, *et al.*: 'Spatial and channel-wise attention in convolutional networks for image captioning'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017, pp. 5659–5667

[12] Pedersoli, M, Lucas, T, Schmid, C, *et al.*: 'Areas of attention for image captioning'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017, pp. 1242–1250

[13] Simonyan, K, Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', arXiv preprint arXiv, September 2014, 1409.1556

[14] He, K, Gkioxari, G, Dollåar, P, *et al.*: 'Mask r-cnn'. Proc. IEEE Int. Conf. Computer Vision (ICCV), Hawaii, USA, 2017, pp. 2961–2969

[15] Box, G. E. P., Tiao, G. C.: '*Bayesian inference in statistical analysis*', vol. **40**, (John Wiley & Sons, Oxford, UK, 2011)

[16] Lin, T.Y., Maire, M, Belongie, S, *et al.*: 'Microsoft coco: common objects in context'. Proc. European Conf. Computer Vision (ECCV), Zurich, Switzerland, September 2014, pp. 740–755

[17] 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. Linguist.*, 2014, **2**, pp. 67–78

[18] Papineni, K., Roukos, S., Ward, T., *et al.*: 'Bleu: a method for automatic evaluation of machine translation'. 40th Annual Meeting on Association for Computational linguistics, Linguistics, Philadelphia, USA, 2002, pp. 311–318

[19] Denkowski, M, Lavie, A.: 'Meteor universal: language specific translation evaluation for any target language'. 9th Workshop on Statistical Machine Translation, Baltimore, USA, 2014, pp. 376–380

[20] Liu, C, Mao, J, Sha, F, *et al.*: 'Attention correctness in neural image captioning'. 31st AAAI Conf. on Artificial Intelligence, San Francisco, USA, February 2017