Explainable Image Caption Generator Using Attention and Bayesian Inference

Seung-Ho Han and Ho-Jin Choi
School of Computing, KAIST
Daejeon, Korea (South)
Email: {seunghohan, hojinc}@kaist.ac.kr

***Abstract*—Image captioning is the task of generating textual descriptions of a given image, requiring techniques of computer vision and natural language processing. Recent models have utilized deep learning techniques to this task to gain performance improvement. However, these models can neither distinguish more important objects than others in a given image, nor explain the reasons why specific words have been selected when generating captions. To overcome these limitations, this paper proposes an explainable image captioning model, which generates a caption by indicating specific objects in a given image and providing visual explanation using them. The model has been evaluated with datasets such as MSCOCO, Flickr8K, and Flickr30K, and some qualitative results are presented to show the effectiveness of the proposed model.**

Keywords—image captioning, objects, visual explanation

I. Introduction

Over the past decades, machine learning(ML) approaches have been applied to a variety of AI fields. Especially, deep learning has shown a great performance in recent years. Computer vision(CV) and natural language processing(NLP) are two research areas that have grown around deep learning methods, such as object detection [1-2] and automatic machine translation [3]. Image captioning, a subfield of CV and NLP, is the task of generating a textural description of a given image. 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.

These existing image captioning models have some limitations. First, the models using encoder–decoder cannot distinguish more important objects than others in a given image, because it generates the caption only using a feature vector for the full input image. Second, the models cannot explain the reasons why specific words have been selected when generating captions. Instead of tackling these issues directly, many developers have tried to only modify hyper-parameters and retrain the model to hope to get better captions, resulting in exhaustive time consumption for model training. We believe that in order to handle these problems properly, an image captioning model should be able to reveal some evidence why the words were generated. In this paper, we propose such a model, so-called "explainable image caption generator", which generates a caption for a given image and explains why specific words have been selected in the generated caption by marking the regions on the image corresponding to the selected words.

This paper is organized as follows. Section 2 provides an overview of related works. Section 3 presents the process and architecture of our proposed model, and section 4 shows the experimental results. Section 5 summarizes the paper.

II. Related Works

A. 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.

B. 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, then trains the generative model using multi-modal recurrent neural network (m-RNN) using image–caption data and pre-calculated scores. Using the trained model, a phrase is generated for an input region. Johnson et al. [8] proposed DenseCap to generate dense captioning (phrase descriptions) for selected regions, using fully convolutional localization networks. The localization layer proposes regions from an input image and extracts their features. Using these features, a RNN language model is trained, which generates short captions for selected regions as the final output.

C. Image Captioning with Attention Mechanism

Neural processes involving attention have been chiefly studied in the computational neuroscience. In the last few years,

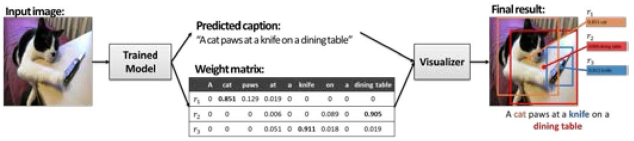
This work was funded by Korea Meteorological Administration Research and Development Program under Grant KMI(2017-00410), and also supported by Korea Electric Power Corporation (Grant number:R18XA05).

978-1-7281-1360-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CSCI46756.2018.00098

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 [9], which showed better performance than previous neural caption generators such as [6] which did not use an attention mechanism. This model [9] also highlights the very image part on which the attention layer focuses when generating each word. As another example, Chen et al. [10] proposed to use multiple attention models for spatial, activation, object etc., and showed better performance than single attention model.

III. Proposed Model

This section describes our proposed model. Our goal is to design an architecture that can generate the caption reflecting the detected objects (i.e., the "regions" syntactically) for a given image, and present the explanation why the caption was generated. To achieve this goal, we propose a model called "explainable image caption generator" as a new approach for explanation. Fig. 1 shows the process of image captioning and visualization.



	A	cat	paws	at	a	knife	on	a	dining table
r_1	0	0.851	0.129	0.019	0	0	0	0	0
r_2	0	0	0	0.006	0	0	0.089	0	0.905
r_3	0	0	0	0.051	0	0.911	0.018	0	0.019

Fig. 1. The process of image captioning and visualization.

A. Image Captioning and Visualization

Assuming that the model training has completed ("how to" will be presented in the next subsection), 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 visualized final result shows several elements: color-coded words in the generated caption, colored region boxes on the image capturing the objects detected, and weight values for the word-region pairs of the same color. 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.

B. The Proposed Model

Fig. 2 shows the architecture of our proposed model, which consists of two modules: (1) generation, and (2) explanation. The generation module generates a caption from the given image using an encoder–decoder architecture. The explanation module generates a weight matrix for all the detected regions (i.e., the objects) in the input image vs. all the words in the generated caption. These two modules also generates loss values, $Loss_g$ and $Loss_e$. Both loss values affect the trainable parameters of the generation module to consider objects. Details of each module is described as follows.

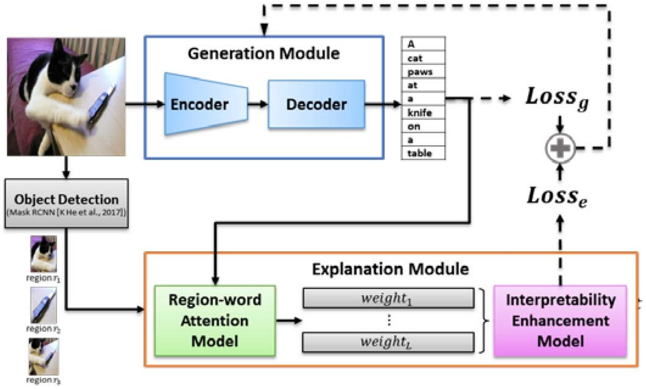


Fig. 2. The architecutre of the proposed model.

1) Generation Module: This module is based on CNN-RNN encoder–decoder framework. The encoder extracts a feature vector for a full image, and the decoder generates the words using the feature vector. For the encoder, we use the VGG-16 model owing to its state-of-the-art performance. To extract the image feature vector, we convert the size of all the images into a fixed size. To implement the decoder model, we use the long short-term memory (LSTM). The decoder model generates the words every LSTM steps using the image feature vector and word distribution. We use a *softmax* function for last layer of LSTM to generate each word. We also use a negative log likelihood loss function to train the encoder–decoder model. As training progresses, the parameters of encoder and decoder are jointly optimized to generate a full caption more accurately. However, this module cannot identify specific parts of the given image. Hence, we designed the explanation module in order for the generation module to consider the important objects that are detected from a given image when generating the caption and providing explanation from the generated caption.

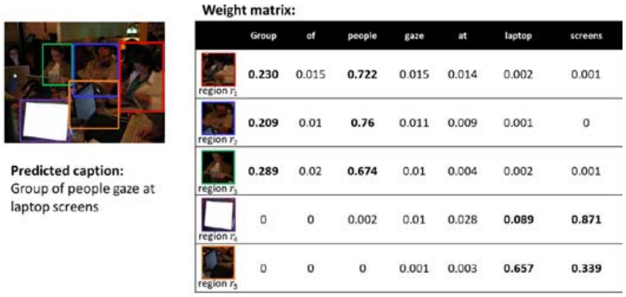
2) Explanation Module: This module has two major roles depending on whether it is in the training or in the testing stage. During training, the explanation module generates $Loss_e$, an image-sentence relevance loss, which digitizes whether the generated caption considers the objects in the input image well. The more the generation module is trained, the better the model can generate a caption considering objects. During testing, the explanation module generates the weight matrix for the regions extracted from the input image and words generated from the generation module 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 module has two components: (1) the region–word attention model, and (2) the interpretability enhancement 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 interpretability enhancement (IE) model generates the image-sentence relevance loss using the weight matrix to assess whether a caption generated from the generation module well-reflects the objects.

IV. Experiments

For experiments, we used three datasets: MSCOCO [11], Flickr8K [12], and Flickr30K [13]. 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 Flickr8K dataset consists of 8,000 images split into 6,000 images for training and 1,000 images for validation and testing. The Flickr30K is an extension of Flickr8K. It consists of 31,783 images with 158,915 crowdsourced captions, and we used it by splitting 29,000 images for training, 1,000 for validation and 1,000 for testing, to make fair comparisons with our baseline paper [7]. Note that each single image in all the datasets comes with five descriptive captions written by human.

A. *Result Analysis: Explanation Module*

We show qualitative results of our region–word attention model in the explanation module. The attention model is provided with some regions (i.e., object areas) detected from the input image and words in the caption generated from the generation module. The output is a weight matrix where each row vector represents the weight values of several candidate words for each region. Fig. 3 illustrates the results, showing the test image, a predicted caption for the image and a weight matrix. In the test image, different region boxes are indicated by borders with different colors. The text below the image represents the predicted caption generated from the generation module. As mentioned earlier, these regions and words are used as an input to the attention model. The weight matrix next to the image is the output of the attention model, where each weight value represents the relevance between the object and the word in the pair.



Predicted caption:
Group of people gaze at laptop screens

Weight matrix:

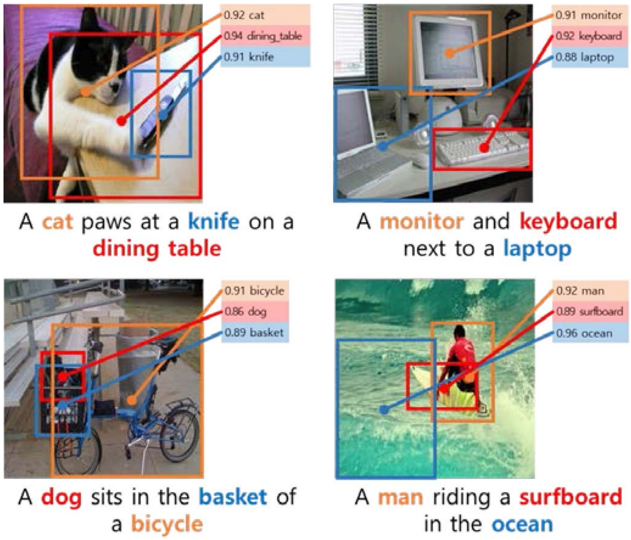
	Group	of	people	gaze	at	laptop	screens
region r_1	**0.230**	0.015	**0.722**	0.015	0.014	0.002	0.001
region r_2	**0.209**	0.01	**0.76**	0.011	0.009	0.001	0
region r_3	**0.289**	0.02	**0.674**	0.01	0.004	0.002	0.001
region r_4	0	0	0.002	0.01	0.028	**0.089**	**0.871**
region r_5	0	0	0	0.001	0.003	**0.657**	**0.339**

Fig. 3. Example of the region-word attetnion model for a given image and predicted caption.

In this example, five regions are detected; three regions, r_1, r_2 and r_3, are labeled as “*person*” and two regions, r_4 and r_5, are labeled as “*screen*” and “*laptop*”. For the regions r_1, r_2 and r_3, the attention model assigns the highest weight value to a word “**people**” (each 0.722, 0.76 and 0.674) because the label of them is a “*person*”. The next highest weight value is assigned to a word “**group**” (each 0.23, 0.209 and 0.289). The reason why the word “**group**” was assigned second highest value is because many ground-truth sentences had included the words “**group**” and “**people**” for the images containing many people in the training dataset. The regions r_4 and r_5 have the similar situation. Both of them are paired to words “**laptop**” and “**screen**”, meaning that ‘laptop’ is close to ‘screen’ in the word embedding. The attention model assigns the high weight values to regions r_4 and r_5. Through the results of the attention model, we can easily check the relations between the regions and words for a given image. Therefore, we can use these relations to generate our final results through a visualizer.

B. *Result Analysis: Entire Model*

In our implementation, the final result for an input image shows a generated caption having colored words, for each of which has a paired region with the same color and an evidence value, as shown in Fig. 4. Each example in this figure shows an image with colored boxes that indicate the detected regions, the generated caption having words colored in the same matching colors as their corresponding regions, and the weight values generated from the attention model.



A cat paws at a knife on a dining table

A monitor and keyboard next to a laptop

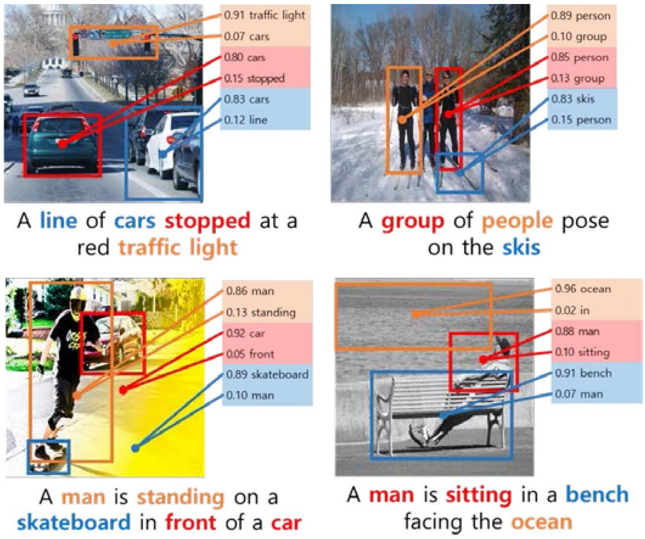
A dog sits in the basket of a bicycle

A man riding a surfboard in the ocean

Fig. 4. Examples of the results generated from our model, which only considered one region-word pair when generating the caption.

Because the generation module was trained using the IE model of the explanation module by picking only one region-word pair for each detected region, only one word can appear for one region in the generated caption. In the first example of Fig. 4, each of the colored words, “**cat**”, “**knife**”, and “**dining table**”, has been generated from the corresponding region, e.g., the word “**cat**” colored orange corresponds to the orange region box encircles the cat with a weight value of 0.92. In this manner, we connect the regions and the specified words, and provide evidence (or explanation) why the words have been selected.

To investigate further the explanation capability of our model, we have diversified the training process using the IE model such that picking two region-word pairs for each detected region, which provides us with an interesting results as illustrated in Fig. 5.



A line of cars stopped at a red traffic light

A group of people pose on the skis

A man is standing on a skateboard in front of a car

A man is sitting in a bench facing the ocean

Fig. 5. Example of final results generated from our model, which considered two region-word pairs for each region when generating the caption.

These examples show that up to two words in the generated caption are paired for each detected region of one image. In the caption "A man is standing on a skateboard in front of a car" generated for the first image in the second row, for instance, the orange-colored words "**man**" and "**standing**" are matched with the orange region box encircles the person, with the weight values 0.86 and 0.13, respectively, representing the first and the second relevant concepts regarding the region. Combined together, this explains the reason why "a man is standing" appears in the generated caption. Comparing to the example of Fig. 4, the results of Fig. 5 include more than just objects but concepts (that is, not only a noun for the object in the given region, but also an adjective describing the state of the concept). In some examples, however, it may not be effective to have two words for one region. In the last result in the second row, for instance, the weight values of "**in**" in the orange box and that of "**man**" in the blue box are too small, hence, not very effective.

V. Conclusion

In this paper, we proposed an explainable image caption generator, a model which generates a caption by considering objects for a given image and provides explanation why the words in the generated caption are selected. To this end, we designed an explanation module that generates an image–sentence relevance loss that influences the training in the generation module. Moreover, the explanation module generates a weight matrix representing the relations for the regions extracted from the given image and words in the generated caption. The weight matrix is used for visualizing the relationship between regions and words. In the qualitative results of our experiments, we demonstrated the advantages of our model in terms of generating descriptive captions and providing explanations for the output. In the future, we plan to improve our model by overcoming some limitations as mentioned previously. In particular, we plan to develop a semantic attention module that can discover more detailed attributes for each region.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2017, pp. 2980–2988.

[2] J. Redmon, S. Divvala, 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 (CVPR),* 2016, pp. 779-788.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[4] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[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 (CVPR),* 2015, pp. 3156–3164.

[7] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[8] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4565–4574.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[10] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," *arXiv preprint arXiv:1611.05594*, 2016.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dolla´r, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[12] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[13] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.