J Inf Process Syst, Vol.15, No.5, pp.1171~1178, October 2019
https://doi.org/10.3745/JIPS.04.0141

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

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Image Understanding for Visual Dialog

Yeongsu Cho* and Incheol Kim**

Abstract

This study proposes a deep neural network model based on an encoder–decoder structure for visual dialogs. Ongoing linguistic understanding of the dialog history and context is important to generate correct answers to questions in visual dialogs followed by questions and answers regarding images. Nevertheless, in many cases, a visual understanding that can identify scenes or object attributes contained in images is beneficial. Hence, in the proposed model, by employing a separate person detector and an attribute recognizer in addition to visual features extracted from the entire input image at the encoding stage using a convolutional neural network, we emphasize attributes, such as gender, age, and dress concept of the people in the corresponding image and use them to generate answers. The results of the experiments conducted using VisDial v0.9, a large benchmark dataset, confirmed that the proposed model performed well.

Keywords

Attribute Recognition, Image Understanding, Visual Dialog

1. Introduction

With the recent advances in computer vision and natural language processing technologies, studies have been actively conducted to solve complex intelligence problems, such as image/video captioning, visual question answering (VQA) [1], and visual dialogs, which require both of the abovementioned technologies. Generally, for VQA, questions and answers regarding an input image are exchanged, and the questions are assumed as mutually independent. However, visual dialogs, an extension of VQA, refer to the continuous exchange of questions and answers regarding one image, and it is assumed that the questions are directly or indirectly interdependent [2-5]. The first Visual Dialog Challenge [2] held in 2017 evaluated agent models that could automatically perform a visual dialog, as in the example illustrated in Fig. 1. In this Visual Dialog Challenge, a dialog proceeds as follows. When a questioner asks a question based on a caption comprising a brief description of an image, a responder answers with an appropriate response to the question by referring to the input image, caption, and dialog history.

Previous studies regarding answerer agent models have primarily focused on the interdependence between questions forming a dialog and the dialog context. For example, [2] presented a method of attention to the previous dialog history that was closely related to the current question. [3] proposed a method using attention memory to focus on visual references common to different questions. Therefore,

Manuscript received June 26, 2019; first revision July 16, 2019; accepted September 4, 2019.
Corresponding Author: Incheol Kim (kic@kyonggi.ac.kr)
* Dept. of Computer Science, Graduate School of Kyonggi University, Suwon, Korea (whdudtn73@gmail.com)
**Dept. of Computer Science, Kyonggi University, Suwon, Korea (kic@kyonggi.ac.kr)

it is important to accurately identify the dialog context based on the dialog history to generate correct answers to questions in a visual dialog. However, in-depth image analysis and understanding for recognizing various attributes of objects or persons in an image are important. Nonetheless, many existing studies on VQA and visual dialogs simply use the visual feature map of the entire image extracted through a convolutional neural network (CNN) to generate answers; few researchers have used the recognition results of objects or persons contained within an image.

This study proposes a deep neural network model based on an encoder–decoder structure for visual dialogs. In the encoding stage of the proposed model, the visual feature map extracted from the entire input image using CNN is used. In addition, rich semantic features extracted by a separate person detector and a person attribute recognizer are used. The person detector detects the region of each person in the image. The person attribute recognizer receives the detected regions of each person as the input and extracts attributes, such as gender, age, and dress concept of the person. In the decoding stage of our model, the most appropriate answer in the answer list is selected based on the fused features obtained from the encoder. We performed various experiments using VisDial v0.9 [2], a large benchmark dataset, to analyze the performance of the proposed model. Subsequently, the experimental results are presented.

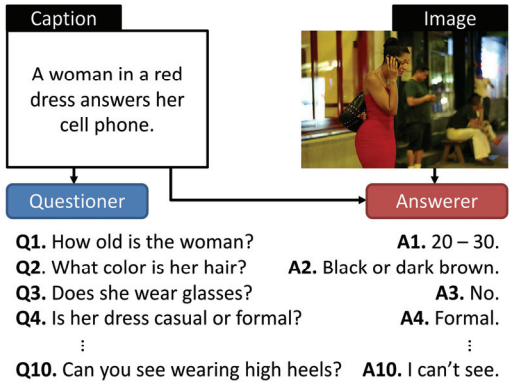

Fig. 1. Example of a visual dialog.

2. Related Work

2.1 Visual Question Answering

Existing studies on VQA [1] have provided a number of technologies required for visual dialogs. In these studies, various attention mechanisms have been developed that focused on regions targeted for questions in an image [6-9]. In particular, [6] proposed a combined bottom-up and top-down attention mechanism. The bottom-up mechanism based on the faster R-CNN proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. [7] proposed a model that used both region attention and language attention. [8] proposed a VQA model that sequentially selects from detected objects and learns the interactions between objects that affect subsequent selections to solve counting questions. Furthermore, [9] presented a model to answer questions for regular objects, e.g., cylinders, spheres, and rectangles, by recognizing their attributes such as shape, color, and material. By contrast, the answerer model for visual dialogs in our study differs vastly

from that of [9], in that it recognizes the attributes of persons in an image and uses them with visual features from the image.

2.2 Visual Dialog

Visual Dialog [2] is a visual dialog environment where free-form questions and answers are exchanged based on a given image; it has been used for the Visual Dialog Challenge since 2017. In this environment, VisDial v0.9, a dataset of visual dialogs between people collected online via the Amazon Mechanical Turk (AMT), is provided. A study on Visual Dialog [3] proposed an answer model employing attention memory. In this model, attentive visual references for each question are stored in an attention memory during a dialog. When similar or related questions are raised, similar visual regions are retrieved from the attention memory to identify potential answers. In general, a discriminative dialog model, which ranks candidate answers given by people, is learned. This provides superior performance over a generative dialog model that is learned to generate answer sentences on its own. However, the discriminative dialog model cannot be applied to an actual dialog. To overcome this limitation, [4] proposed a method of knowledge transfer from a discriminative dialog model to a generative dialog model. [5] proposed a new dialog model that combined reinforcement learning with generative adversarial networks to generate more human-like answers to questions. However, a dialog model that generates correct answers through deeper understanding of images based on persons and objects in images has not been established.

3. Visual Dialog Model

In this study, we designed an answerer model that selects the most appropriate answer in the candidate answer list for a given question with an encoder–decoder structure. As depicted in Fig. 1, an answerer is given an image and a caption regarding the image. Each time a questioner asks a question regarding the image, a dialog continues when an appropriate answer is generated by the answerer. Thus, a given image (Image I), a question from the current round of questions (Question Q_t), and a dialog history H_t = <(Caption C), (Q_1, A_1), ..., (Q_{t-1}, A_{t-1})> up to the previous round including the caption, are provided as the input to the answerer encoder. In the encoder, the input features are extracted through a separate network. By combining the extracted features, the final encoded features are obtained. The proposed model is a discriminative dialog model that learns to rank the suggested candidate answers. Therefore, as the input, the decoder of the proposed model receives the output e_t of the encoder and a list of k candidate answers ($Answer\_List_t$) corresponding to the current question. The decoder chooses the most appropriate answer from the candidate answer list.

3.1 Encoder

The proposed encoder extracts the linguistic feature vector l_t from the current question Q_t and the dialog history H_t. Along with the visual feature vector extracted from the input image, the linguistic feature vector is used to generate the final encoded feature vector. The structure of the proposed encoder is illustrated in Fig. 2. The current question Q_t and the dialog history H_t = <(Caption C), (Q_1, A_1), ..., (Q_{t-1}, A_{t-1})> are natural language texts. Therefore, features for the image caption C, the question and answer pair (Q_i, A_i) of each round, and the current question Q_t are encoded through the layers of Word

Embedding and long short-term memory (LSTM) [4], which is a recurrent neural network. The next phase focuses on the question–answer pair (Q_i, A_i) most closely related to the current question Q_t in the previous dialog history H_t. The correlation between the feature vector of the current question Q_t and the feature vector of each question–answer pair (Q_i, A_i) constituting the dialog history H_t, is calculated by an inner product operation. The calculated degree of correlation with the current question Q_t is used as the weight for each question–answer pair (Q_i, A_i) constituting the dialog history H_t. By calculating the weighted sum of the question–answer pairs (Q_i, A_i) based on these weights, the final feature vector for the dialog history H_t is obtained. The feature vector of the dialog history obtained, H_t is concatenated with the feature vector of the current question, Q_t. Subsequently, a linguistic feature vector l_t is generated through a fully connected layer.

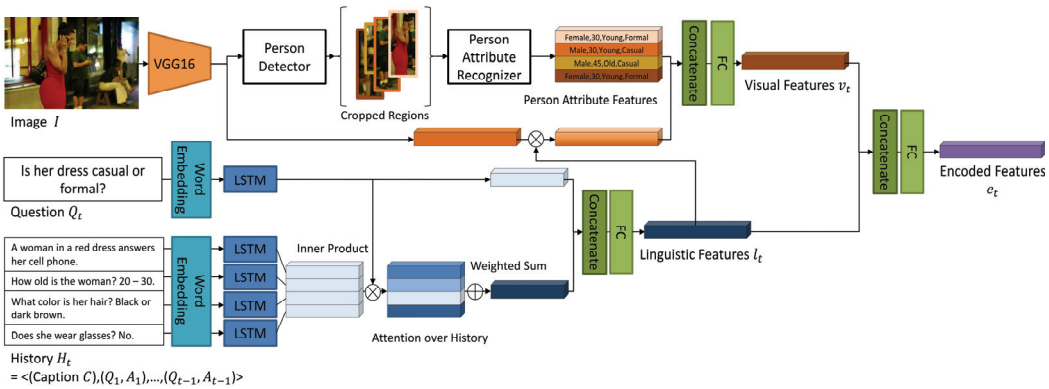

Fig. 2. Encoder structure.

The encoder extracts the attributes of the persons contained in the input images. Additionally, the encoder focuses on the most relevant areas in the entire image using a linguistic feature vector. First, it extracts visual features for the entire image from an input image I through VGG16 [2], which is a typical CNN. Subsequently, it uses YOLO v3 as a person detector, having learned in advance to detect persons in the MS COCO dataset [2]. It detects the region of each person in the visual feature map using a person detector. The cropped regions obtained from the person detection stage then continues to the stage of person attribute recognition. DeepMAR trained with PETA, the largest challenging pedestrian attribute dataset, is used as a person attribute recognizer. In this study, we only extracted attributes of gender, age, and dress concept for each person by modifying DeepMAR. Here, the gender attribute can have one of two values, "female" or "male," and the age attribute can have one of four values, i.e., "less than 30," "30 to 44," "45 to 59," and "60 and above." Additionally, the age attribute can have either of two values, "young" for 30 years old or less, or "old" for all others. Finally, the dress attribute can only have one of two values, "casual" or "formal." The encoder uses an attention mechanism to obtain the most relevant areas in the input image to the current question Q_t and dialog history H_t. The correlation between the visual feature vector and the linguistic feature vector is calculated by an inner product operation. Subsequently, the calculated values are used as weights via the softmax layer. By applying these weights to the entire visual feature, the focused visual feature vector is obtained. Subsequently, the focused visual feature vector and the person attributes vector are combined into the final visual feature vector v_t through a simple concatenation and a fully connected layer. In the final stage, the visual feature vector v_t and the

linguistic feature vector l_t are combined into the final encoded vector e_t through a simple concatenation and a fully connected layer.

3.2 Decoder

The proposed discriminative decoder selects the most appropriate answer in the answer list ($Answer\_List_t$) based on the fused feature information e_t obtained from the encoder. The structure of the proposed identification decoder is shown in Fig. 3. The discriminative decoder encodes each candidate answer $a_{t,i}$ in the answer list ($Answer\_List_t$), received as an input via Word Embedding and then LSTM. Subsequently, the decoder computes a dot product with each encoded answer feature and the encoder output e_t to obtain the correlation score between them. Each dot product value is converted into a candidate answer score $s_{t,i}$ when passed through softmax and is stored in the score list ($Score\_List_t$). The decoder is learned to minimize cross entropy errors using the index of correct answers and the score list. When using the learned decoder to generate an answer to a given question, the answer with the highest score in the score list becomes the output.

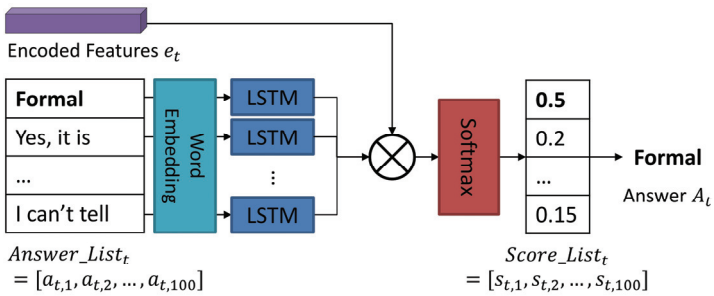

Fig. 3. Decoder structure.

4. Implementation and Experiments

4.1 Implementation

The VisDial v0.9 benchmark dataset was used in this study to test the performance of the proposed model. This dataset comprises authentic dialog data collected using the AMT. VisDial consists of 10 question–answer pairs per image. MS COCO data were used as the image data given to a dialog; 18,191 of the data were used as training data and the remaining 8,898 were used as test images. For implementation and experiments, we used a Geforce GTX 1080 Ti GPU, and the model was implemented using PyTorch, a Python deep-learning library, under the Ubuntu 16.04 LTS operating system. Our model used two-layer LSTMs. The dimension of the word embeddings is 300. For training the model, the batch size was set to 12, and the iteration count was set to 20.

4.2 Experiments

The intent of the first experiment was to prove that the encoder with the added attributes of a person can improve the performance. We evaluated whether each individual attribute could improve the

performance while adding attributes to the encoder of the proposed model. In Table 1, "VF" represents the case where only the visual features of an image were used, without using the attributes of a person. Meanwhile, "VF+Age" refers to the case where the age attribute is added to the VF of the image, and "VF+Age+Gender" and "VF+Age+Clothes" refer to cases where the gender and dress attributes are added separately to "VF+Age." Finally, "VF+All" represents the case where not only the visual features of an image but also the gender, age, and dress attributes are used together, as in the proposed model.

We used the three evaluation measures presented in [2]. First, the mean reciprocal rank (MRR) is the average reciprocal number of the rank of correct answers in the answer list predicted by the model. The MRR is expressed as follows:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}, \tag{1}$$

where Q refers to the entire question set and $rank_i$ refers to the rank of the correct answer for the i^{th} question. Next, Recall@k refers to the probability that a correct answer exists to the question among the k top-ranked answers in the answer rank list predicted by the model. Finally, Mean refers to the average rank of the correct answers in the answer rank list predicted by the model. For example, if each answer to five questions has a rank of [6, 8, 10, 1, 3], the mean is 5.6, which is the average of the ranks.

Table 1. Performance comparison of different feature sets

Features	MRR	R@1	R@5	R@10	Mean
VF	0.6409	52.65	76.99	83.66	6.79
VF+Age	0.6576	54.78	78.18	84.43	6.91
VF+Age+Clothes	0.6595	55.60	77.54	83.53	7.01
VF+Age+Gender	0.6657	56.23	78.24	84.17	6.87
VF+ALL	0.6691	56.35	78.93	85.83	6.36

As shown in Table 1, the cases where the attributes of a person are added show better overall performances compared with the case of "VF," where the attributes of a person are not used. Additionally, it shows that adding more attributes results in a better performance. This means that, to a certain degree, each attribute contributes to the performance improvement. When comparing "VF+Age+Clothes" with "VF+Age+Gender," it is clear that the gender attribute is more effective than the clothing attribute. We confirmed that the performance improvement differs for certain attributes. The proposed model, "VF+All," which uses all the attributes of the recognized person, shows the best performance in all evaluations.

In the second experiment, the proposed model is compared with other recent state-of-the-art models to prove its superior performance. Table 2 presents the results of the performance comparison. The models used for comparison were the late fusion (LF), hierarchical recurrent encoder (HRE), and memory network (MN) of [2], attention memory (AMEM) of [3], and history-conditioned image attentive encoder (HCIAE) of [4]. LF [2] extracts features from an image, question, and dialog history through individual module networks and then combines them together. The HRE [2] applies a hierarchical recurrent encoder to encode the dialog history. The MN [2] treats each question–answer pair of each round as a fact in the memory bank and computes the attention over the fact using a question

and an image to obtain the most related fact to answer a question. The AMEM [3] stores the previous (visual attention, key) pairs in the attention memory bank and retrieves the most relevant visual attention for the question to resolve the current reference in the visual dialog. The HCIAE [4] attends to the history using questions and attends to the image using questions and attended history to obtain the final encoded features.

Table 2. Performance comparison with the state-of-the-art models

Model	MRR	R@1	R@5	R@10	Mean
LF [2]	0.5807	43.82	74.68	84.07	5.78
HRE [2]	0.5868	44.82	74.81	84.36	5.66
MN [2]	0.5965	45.55	76.22	85.37	5.46
AMEM [3]	0.6160	47.74	78.04	86.84	4.99
HCIAE [4]	0.6222	48.48	78.75	87.59	4.81
Proposed model	0.6691	56.35	78.93	85.83	6.36

As presented in Table 2, the proposed model exhibits better performance than the other models in terms of the MRR, R@1, and R@5. For the MRR, the proposed model shows performance improvements of 15.22%, 14.03%, 12.17%, 8.62%, and 7.54% over LF, HRE, MN, AMEM, and HCIAE, respectively. Furthermore, for R@1 and R@5, which require a more stringent accuracy than R@10, the proposed model exhibits a significantly better performance than the other models, proving the excellent performance of the proposed model. However, the proposed model shows similar or slightly degraded performance than the other models in terms of R@10 and Mean.

5. Conclusions

Existing studies on visual dialogs merely use the visual features of the entire image extracted through a CNN to generate answers. No researcher has emphasized the attributes of objects or persons contained in images to subsequently use them as features. We herein proposed a deep neural network model based on an encoder–decoder structure for visual dialogs. At the encoding stage of the proposed model, we emphasized attributes, such as gender, age, and dress concept of the persons contained in the images and subsequently used them in generating answers by employing a separate person detector and an attribute recognizer. Through various experiments using VisDial v0.9, a large benchmark dataset, we confirmed the effective performance of the proposed model. In the future, we intend to conduct research to improve the accuracy of the attribute recognizer employed in the proposed model and develop effective attention mechanisms between questions and images.

Acknowledgement

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

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4-31, 2017.

[2] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 326-335.

[3] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3719-3729, 2017.

[4] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, "Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model," *Advances in Neural Information Processing Systems*, vol. 30, pp. 314-324, 2017.

[5] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? Reasoned visual dialog generation through adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 6106-6115.

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 6077-6086.

[7] L. Peng, Y. Yang, Y. Bin, N. Xie, F. Shen, Y. Ji, and X. Xu, "Word-to-region attention network for visual question answering," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3843-3858, 2019.

[8] A. Trott, C. Xiong, and R. Socher, "Interpretable counting for visual question answering," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[9] M. T. Desta, L. Chen, and T. Kornuta, "Object-based reasoning in VQA," in *Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 2018, pp. 1814-1823.

Yeongsu Cho https://orcid.org/0000-0002-8020-6310

She received her B.S. degree in Computer Science from Kyonggi University in 2018. She is currently an M.S. student of the Department of Computer Science, Kyonggi University, Korea. Her current research interests include machine learning, computer vision, and intelligent robotic systems.

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

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