Computer Science and Information Systems 18(1):155–168 https://doi.org/10.2298/CSIS200515038G

Double-Layer Affective Visual Question Answering Network

Zihan Guo[1], Dezhi Han[1], Francisco Isidro Massetto[2], and Kuan-Ching Li[3]*

[1] College of Information Engineering, Shanghai Maritime University
Shanghai 201306, China
guo_zihan11@163.com, dzhan@shmtu.edu.cn
[2] Center for Cognition and Complex Systems, Universidade Federal do ABC (UFABC)
Santo André, SP 09210-580, Brazil
francisco.massetto@ufabc.edu.br
[3] Dept. of Computer Science and Information Engineering, Providence University
Taichung 43301, Taiwan
kuancli@pu.edu.tw

Abstract. Visual Question Answering (VQA) has attracted much attention recently in both natural language processing and computer vision communities, as it offers insight into the relationships between two relevant sources of information. Tremendous advances are seen in the field of VQA due to the success of deep learning. Based upon advances and improvements, the Affective Visual Question Answering Network (AVQAN) enriches the understanding and analysis of VQA models by making use of the emotional information contained in the images to produce sensitive answers, while maintaining the same level of accuracy as ordinary VQA baseline models. It is a reasonably new task to integrate the emotional information contained in the images into VQA. However, it is challenging to separate question-guided-attention from mood-guided-attention due to the concatenation of the question words and the mood labels in AVQAN. Also, it is believed that this type of concatenation is harmful to the performance of the model. To mitigate such an effect, we propose the Double-Layer Affective Visual Question Answering Network (DAVQAN) that divides the task of generating emotional answers in VQA into two simpler subtasks: the generation of non-emotional responses and the production of mood labels, and two independent layers are utilized to tackle these subtasks. Comparative experimentation conducted on a preprocessed dataset to performance comparison shows that the overall performance of DAVQAN is 7.6% higher than AVQAN, demonstrating the effectiveness of the proposed model. We also introduce more advanced word embedding method and more fine-grained image feature extractor into AVQAN and DAVQAN to further improve their performance and obtain better results than their original models, which proves that VQA integrated with affective computing can improve the performance of the whole model by improving these two modules just like the general VQA.

Keywords. deep learning, natural language processing, computer vision, visual question answering, affective computing.

* Corresponding author

1. Introduction

In recent years, multimodal learning for natural language processing (NLP) and Computer Vision (CV) has gained broad interest, such as Visual Question Answering (VQA) [1], image captioning [41] and image-text matching [10], among several others [24]. Compared to other multimodal learning tasks, VQA is more challenging, since it requires a fine-grained understanding of both textual questions and visual images, and it may also involve complex reasoning and require common sense knowledge to answer the questions correctly. Therefore, VQA is regarded as a test of the deep visual and textual understanding ability of a model, as well as a benchmark for general artificial intelligence (AI). An instance of VQA consists of typical tasks that connect an image and a related question, so the task of the machine is to produce the correct answer. There are many potential applications for VQA, such as a personal assistant or robotics designed to assist individuals with physical disabilities.

In early VQA models, the conventional approach is to train a deep neural network with supervision, which maps the given question and the given image to a relative scoring of candidate answers. Specifically, the input question is first tokenized into words, so then the model utilizes the word embedding method to transform the terms into single vectors. Next, the model inputs the word vectors into a Recurrent Neural Network (RNN) to obtain the question features and inputs the given image into a Convolutional Neural Network (CNN) pre-trained on object recognition to capture the image features. Finally, the model fuses the question features and the image features through linear pooling (such as element-wise multiplication) and then feeds the joint embedding into a classification layer to predict the correct answer. With the emergence of advanced word embedding methods, fine-grained feature extractors, cognitive fusion mechanisms and various attention mechanisms, the performance of VQA models is also improved.

It is noteworthy that most of the existing VQA models do not further understand nor analyze the emotional information contained in the input images. Part of the reason is due to the fact that, there is no VQA dataset that includes rich emotional information to the images it contains so far. Till recently, the Affective Visual Question Answering Network (AVQAN) [34] enriches the model's understanding and analysis of VQA by making use of the emotional information contained in the images to produce sensitive answers, while maintaining the same level of accuracy as ordinary VQA baseline models. It is a reasonably new task to integrate the emotional information contained in the images into VQA.

However, it is challenging to separate question-guided-attention from mood-guided-attention in AVQAN, due to the concatenation of question words and mood labels. It is believed that this type of concatenation is hazardous to the performance of the model. To mitigate this effect, we propose the Double-Layer Affective Visual Question Answering Network (DAVQAN), which divides the task of generating emotional answers in VQA into two relatively simple subtasks, i.e., the generation of non-emotional responses and the production of mood labels, and utilizing two independent layers to tackle the two subtasks respectively. In such studies, the emotional information contained in the images refers to human facial expressions. Since there is no publicly available dataset suitable for VQA integrated with affective computing, we use the same method as AVQAN to construct a preprocessed dataset to complete the proposed research. We conduct a comparative experiment on the preprocessed dataset to compare the performance of AVQAN and

DAVQAN, and the experimental results show that the overall performance of DAVQAN is 7.6% higher than that of AVQAN, showing the effectiveness of the proposed model. We also introduce more advanced word embedding method and more fine-grained image feature extractor into AVQAN and DAVQAN to further improve their performance and obtain better results than their original models, what shows that VQA integrated with affective computing can improve the performance of the entire model by improving these two modules just like the general VQA.

The remainder of this article is organized as follows. Section 2 reviews the works related to VQA, while in section 3 is provided the details of DAVQAN. Next, details on how we construct the preprocessed dataset for the experiments and experimental evaluation are presented in Section 4, and finally, we present the conclusions and future directions of this work in Section 5.

2. Related work

Text-based Question Answering. Text-based question answering is a longstanding problem that has been studied for decades in natural language processing. The model needs to fully understand the textual questions and requires a wide range of knowledge to answer the questions correctly [26] [23]. The early text-based question answering system [39] uses information retrieval to find out the text containing the answer as the output of the model. Recently, advanced methods, e.g. [3], have improved the accuracy of their answers by constructing large-scale knowledge bases. Through all the efforts, text-based question answering has been successfully applied to search engines, mobile devices, and other fields. Various methods and models for text-based question answering inspire VQA techniques. Nevertheless, unlike text-based question answering, VQA is naturally grounded in images – requiring the understanding of both visual images and textual questions. The information contained in the visual images is more abundant and noisier than that contained in the textual questions. Therefore, VQA is more challenging to deal with than text-based question answering. Meanwhile, the interactions between the visual images and the textual questions are also essential to VQA. Furthermore, the questions are generated by humans, making the need for complex reasoning and common sense knowledge more essential.

Describing Visual Content. For many years, many researches have been devoted to the study of joint learning which combines the visual and textual information [4], [5], [30], [32], [37], [46], [40], [44]. Related to VQA are the tasks of image tagging [21], [15], video captioning [33], [13] and image captioning [22], [8], [27], [41], where the models are used to generate sentences or words to describe visual content. Automatically describing the content of an image is a fundamental problem of artificial intelligence [16]. In the early stages, the researches on describing visual content mainly included the object classification task and the task of assigning descriptions. While these tasks require both semantic and visual knowledge, captions can often be non-specific. Even the more advanced methods and models for generating generic image captions are of little use for VQA, since the questions in VQA require detailed specific information about the images. Therefore, compared with captioning tasks, VQA is more complex, more interactive, and has a broader range of applications.

Visual Question Answering. In the past few years, VQA has attracted more and more attention. The first VQA dataset developed as a benchmark is Data Set for Question Answering on Real World (DAQUAR) [25]. With the continuous development, the most popular modern datasets use images sourced from Microsoft Common Objects in Context (COCO) [38], a dataset initially designed for image recognition. Those images constitute a diverse collection of photographs. Some of the latest versions of these datasets, such as VQA v2.0 [11], have been proposed to address issues of dataset biases and other issues. Based on these datasets, many models have been proposed to deal with VQA tasks. Most of these models learn the joint embedding of the image features and the question features and then input them into a classification layer to predict the correct answer.

From the above description, we can see that in most VQA models, the first step is to use the word embedding method to transform the question words into single vectors. Initially, common word embedding methods included the one-hot representation of words and the GloVe word embeddings [28] pre-trained on a large-scale corpus. ELMo [29], a later proposed method, improves the performance of word embedding methods by concatenating the left-to-right and the right-to-left word features extracted from the text. However, models like ELMo are feature-based and not profoundly bidirectional. At present, the more advanced method BERT [7] can pre-train a deep bidirectional Transformer and can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of natural languages processing tasks, such as word embedding and sentence classification.

After transforming the question words into single vectors, the VQA models need to extract the question features and the image features. The Long Short-Term Memory (LSTM) [14] and the Gated Recurrent Unit [17] are the most common methods to extract question features. And most of the original VQA models use the VGGNet [36] to extract image features. Now the more advanced methods for extracting image features are the ResNet [18] and the bottom-up attention network [2] derived from Faster R-CNN [35]. The question features and the image features obtained by early VQA models are at the global-level and contain noisy information. In many cases, keywords of the questions and the local areas of the images are the key to answer the questions correctly. As a result, various attention mechanisms have been proposed and have become an integral part of VQA models (e.g., [2]). The core idea of attention mechanism is to assign different weights to local features so that the model can focus on the essential local features rather than the global features. Furthermore, the multimodal feature fusion mechanisms of question features and image features are also fundamental to VQA models because of the requirements of the models for understanding and analyzing the content of the input questions and the input images and the relationships between them. The element-wise addition and the element-wise multiplication are the earliest multimodal feature fusion mechanisms used for VQA. To obtain higher-level interactions between question features and image features, several methods based on bilinear pooling have been proposed, such as MCB [9]. With the development of the above technologies, the performance of VQA models also improved.

3. Double-Layer Affective Visual Question Answering Network

Although AVQAN enriches the model's understanding and analysis of VQA, it is difficult for AVQAN to separate question-guided-attention from mood-guided-attention due to the concatenation of the question words and the mood labels. Different from AVQAN, DAVQAN divides the task of generating emotional answers into two relatively simple subtasks, i.e., the generation of non-emotional answers and the generation of mood labels, and uses two independent layers to tackle the two subtasks respectively. The non-emotional layer takes the images, and the questions as input to predict non-emotional answers, and the emotional layer deals with the emotional information contained in the input images to predict mood labels for the images. Finally, we combine the non-emotional answers and the mood labels to compose the emotional answers. In this section, we first introduce the non-emotional layer. The emotional layer will be detailed in the second part.

3.1. Non-emotional layer

In the early stages, the image features used in VQA models were global features and contained irrelevant and noisy information. In many cases, the local areas of the image are the key to answer the question correctly. Thus, the attention mechanisms based on visual attention were proposed and have become an integral part of VQA models. With the development of attention mechanisms, researchers have successfully proposed the co-attention mechanisms [6], [45] that can focus on both the keywords of the questions and the local areas of the images to improve the performance of VQA models. For a fair comparison, DAVQAN uses the same attention mechanism as AVQAN, that is, we use the input questions to guide the model to focus on the local areas of the input images.

We introduce the spatial attention [12], [42] into the standard LSTM to construct our non-emotional layer. The non-emotional layer takes the images and the questions as input to predict the non-emotional answers. It learns to attend to the pertinent regions of the input image as it reads the input question tokens in a sequence. Specifically, the input textual question $Q = (q_1, q_2, \ldots, q_n)$ is first tokenized into words and these words are then transformed into one-hot representations by function $OH(\cdot)$. And the input visual image I is represented as a set of regional image features extracted from a pre-trained CNN model. Now there are many advanced image feature extractors such as the ResNet [18] and the bottom-up attention network [2] derived from Faster R-CNN [35]. For fair comparison, we choose the same image feature extractor as AVQAN, i.e., the VGGNet [36]. In the experimental part, we also replace the VGGNet with ResNet to extract more fine-grained image features to further improve the performance of the models.

The embeddings of the image and the question tokens can be given as follows:

$$v_0 = W_i[F(I)] + b_i\,. \tag{1}$$

$$v_i = W_w[OH(t_i)], i = 1, ..., n\,. \tag{2}$$

where $F(\cdot)$ represents the CNN extractor which transforms the visual image I from the pixel space to a 4096-dimensional feature representation. The W_i matrix and the W_w matrix are used to embed the image feature and the question word embeddings into the same dimension. Thus, we can concatenate the image feature and the question word embeddings and input them into the LSTM model one by one to infuse our attention mechanism. In AVQAN, the embedding of the mood label is also added to the concatenation. We think that in this kind of concatenation, the question-guided-attention and the mood-guided-attention will interfere with each other. The update rules of our non-emotional layer can be defined as follows:

$$\mathrm{i}_t = \sigma(W_{vi}v_t + W_{hi}\mathrm{h}_{t-1} + W_{ri}\mathrm{r}_t + b_i)\,. \tag{3}$$

$$\mathrm{f}_t = \sigma(W_{vf}v_t + W_{hf}\mathrm{h}_{t-1} + W_{rf}\mathrm{r}_t + b_f)\,. \tag{4}$$

$$\mathrm{o}_t = \sigma(W_{vo}v_t + W_{ho}\mathrm{h}_{t-1} + W_{ro}\mathrm{r}_t + b_o)\,. \tag{5}$$

$$\mathrm{g}_t = \tanh(W_{vg}v_t + W_{hg}\mathrm{h}_{t-1} + W_{rg}\mathrm{r}_t + b_g)\,. \tag{6}$$

$$\mathrm{c}_t = \mathrm{f}_t \odot \mathrm{c}_{t-1} + \mathrm{i}_t \odot \mathrm{g}_t\,. \tag{7}$$

$$\mathrm{h}_t = \mathrm{o}_t \odot \tanh(c_t)\,. \tag{8}$$

where $\sigma(\cdot)$ represents the sigmoid function and $\bigodot$ is the element-wise multiplication operator. The convolutional features and the previous hidden state determine the attention term $\mathbf{r}_t$, which is the weighted average of the convolutional features and can be calculated by the following formula:

$$\mathrm{e}_t = w_a^T \tanh(W_{he}\mathrm{h}_{t-1} + W_{ce}C(I)) + b_a\,. \tag{9}$$

$$\mathrm{a}_t = \mathrm{softmax}(\mathrm{e}_t)\,. \tag{10}$$

$$\mathrm{r}_t = \mathrm{a}_t^T C(I)\,. \tag{11}$$

where the pre-trained model VGGNet extracts the 14×14 512-dimensional convolutional image features which are represented by *C(I)*, $\mathbf{e}_t$ represents the embedding of the previous hidden state $\mathbf{h}_{t-1}$, and $\mathbf{a}_t$ stands for a 196-dimensional vector of the image attention weights. The dimension of the regional image features is 512 and each image has 14×14 regions. All the weight matrices *W*s, biases *b*s and the attention terms in our non-emotional layer are learnable parameters. Finally, we relay the last LSTM hidden state to the Softmax classifier to predict the non-emotional answers. Figure 1 shows the structure of our non-emotional layer.

3.2. Emotional layer

Recent works have studied on utilizing CNN for visual attribute detection. In this paper, we follow the study from [31] to build our emotional layer. The emotional layer takes the visual images as input to predict mood labels for the images. In our settings, there are two tasks: the prediction of non-emotional answers and the prediction of mood labels. Both of these tasks share the same lower layers of the pre-trained CNN model VGGNet. The pre-trained CNN model VGGNet takes a square pixel RGB image as input and is composed of five successive convolutional layers C1. . . C5. After C5, there are three fully connected layers FC6. . . FC8. These three fully connected layers compute $\mathbf{Y}_6$=$\sigma(W_6\mathbf{Y}_5 + B_6)$, $\mathbf{Y}_7$=$\sigma(W_7\mathbf{Y}_6 + B_7)$ and $\mathbf{Y}_8$=$\psi(W_8\mathbf{Y}_7 + B_8)$, where $\mathbf{Y}_k$ denotes the output of the *k*-th layer, W_k, B_k are the learnable parameters of the *k*-th layer, and $\sigma(\mathbf{X})[i]$=max$(0, \mathbf{X}[i])$ and $\psi(\mathbf{X})[i]$=$e^{\mathbf{X}[i]}/\Sigma_j e^{\mathbf{X}[j]}$ are the "ReLU" and "SoftMax" non-linear activation functions.

Although the emotional layer and the pre-trained CNN model VGGNet are both designed to tackle the image classification task, the object labels of the two are quite different. To solve this problem, we remove the output layer FC8 of the VGGNet and add an adaptation layer formed by two fully connected layers FCA and FCB. FCA and FCB take

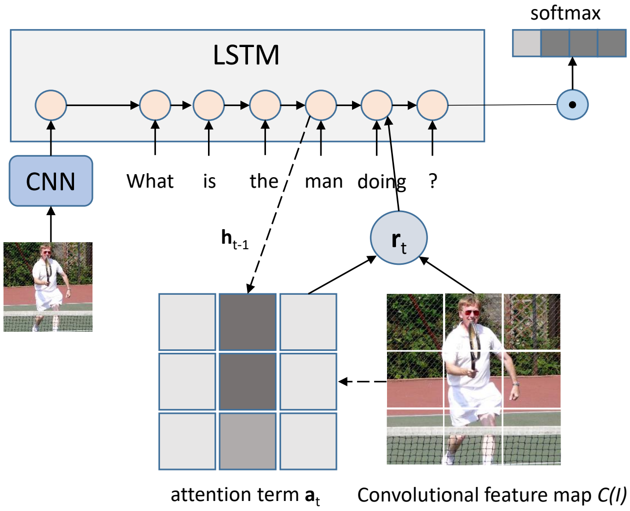

Fig. 1. The structure of the non-emotional layer

the output vector $\mathbf{Y}_7$ of the layer FC7 as input to predict a mood label for the given image. The calculation formula is as follows:

$$\mathrm{Y}_a = \sigma(W_a\mathrm{Y}_7 + B_a)\,. \tag{12}$$

$$\mathrm{Y}_b = \psi(W_b\mathrm{Y}_\mathrm{a} + B_b)\,. \tag{13}$$

where W_a, B_a, W_b, B_b are learnable parameters. In our emotional layer, FC6 and FC7 have the same size 4096, FCA has size 2048 and FCB has a size equal to the number of mood categories.

The layers C1. . . C5, FC6, and FC7 are pre-trained on the ImageNet and then transferred to our mood classification task and kept fixed. The two fully connected layers FCA and FCB are trained on the preprocessed dataset. Figure 2 shows the architecture of the emotional layer.

4. Experiments and results

In this section, we will describe how we construct the preprocessed dataset for our experiments and perform the experimental evaluation.

4.1. The preprocessed dataset

Since there is no publicly available dataset suitable for VQA integrated with affective computing, we use the same method as AVQAN to construct a preprocessed dataset, which is composed of images of people, questions, answers and mood labels, based on the

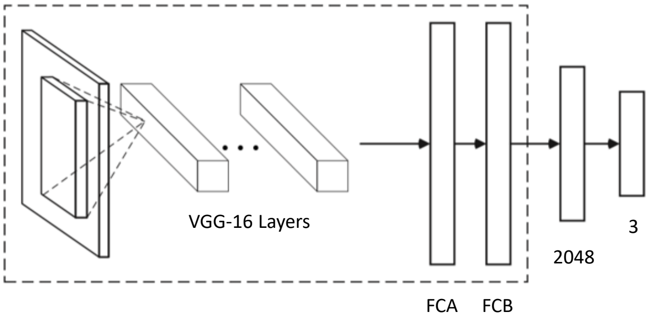

Fig. 2. Deep architecture of the emotional layer

Visual7w dataset [43] to complete our research. The Visual7w dataset is a subset of the Visual Genome QA dataset [20], which is one of the largest datasets designed for VQA with 1.7 million question/answer pairs. Besides, the Visual7w dataset uses the seven questions (What, Where, When, Who, Why, How and Which) to systematically check the visual and textual comprehension capabilities of a model. Note that, the 7th Which question category is used to extend existing VQA setups to accommodate visual answers, which is irrelevant to our study.

Specifically, we remove the images that are irrelevant to our task from the Visual7w dataset, leaving only the images bearing at least one person, and label each image with a mood label. It is worth noting that the Visual7w dataset is not a dataset dedicated to mood classification tasks, and many images of it contain little or no emotional information. Thus, we only obtain a limited set of samples. Considering that the corpus of the question words is tiny, we set the questions in our preprocessed dataset relatively simple to prevent the accuracy of the models from being too low. The 3 mood labels used are happy, surprised, and neutral, for there are too few samples of other mood labels such as sad. Among the total number of instances in the preprocessed dataset, 50% for training, 20% for validation, and 30% for testing. The ratios remain as they are in the AVQAN paper to ensure a fair comparison.

4.2. Experiment setup

During the experiment, the non-emotional layer takes the visual images, and the textual questions as input to predict non-emotional answers, and the emotional layer deals with the emotional information contained in the input images to predict mood labels for the images. If the mood label of an image is neutral, the emotional aspect is ignored in the answer. We use backpropagation to train our model and choose cross-entropy as our loss function. During validating, we use the validating split of the preprocessed dataset for hyper-parameter selection and early stopping. During testing, the model takes the visual images and textual questions as input, and we say the model is correct on a question if it manages to output the correct mood label and the correct non-emotional answer. The dimensions of the LSTM gates and memory cells are 512 in all the experiments, and the model is trained with Adam update rule [19]. In this paper, we evaluate the generated answers in the open-ended setting. An alternative method to evaluate is to let the model pick the correct mood label and the correct non-emotional answer among the candidates.

4.3. Answer categories

The answers generated by our proposed DAVQAN model can be classified as partially wrong (A: having a wrong mood but the rest of the answer is correct or B: having a correct mood but the rest of the answer is wrong), C: completely wrong, or D: completely correct. Figure 3 shows several examples of the four categories and Table 1 shows the accuracy of the four categories during testing.

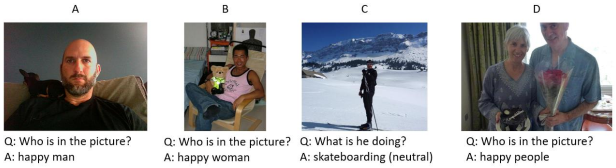
A
B
C
D
Q: Who is in the picture?
A: happy man
Q: Who is in the picture?
A: happy woman
Q: What is he doing?
A: skateboarding (neutral)
Q: Who is in the picture?
A: happy people

Fig. 3. The four answer categories from DAVQAN

Indeed, visual attribute detection is one of the most challenging problems in computer vision. As shown in Table 1, the performance of the emotional layer is not satisfactory. Although there are only three types of mood labels, the accuracy can only reach 79.89%. We think the reason that limits the performance of the emotional layer is that the preprocessed dataset is not large enough. In future studies, we will construct a larger and more suitable dataset for VQA integrated with affective computing to give full play to the advantages of deep learning.

4.4. Comparison with original AVQAN

Table 1. Analyzing the overall percentage of answers in each category for the DAVQAN model

category	accuracy
A	12.50%
B	24.46%
C	7.61%
D	55.43%

Nelson et al. [34] indicate that the integration of affective computing in AVQAN has no significant impact on the performance of ordinary VQA baseline models but rather enriches the model's understanding and analysis of images. In AVQAN, however, it is difficult to separate question-guided-attention from mood-guided-attention due to the concatenation of the question words and the mood labels. To solve this problem, we propose DAVQAN and conduct a comparative experiment on our preprocessed dataset to compare the performance of AVQAN and DAVQAN. To carry out the comparative experiment, we set the emotional answers as supervision for AVQAN, and the non-emotional answers and the mood labels as supervision for DAVQAN. To compare their structures fairly, the two models not only use the same word embedding method, the same feature extractors and the same attention mechanism, but also have the same parameter settings. Table 2 shows the results of AVQAN and DAVQAN on our preprocessed dataset. As observed in Table 2, the overall performance of DAVQAN is 7.6% higher than AVQAN, which shows the effectiveness of the proposed model. Due to the size limitation of the preprocessed dataset, the corpus of question words is very limited. The question-guided-attention is not effective enough to help the model answer questions correctly. Thus, we only count the overall performance of the two models, and the accuracy of AVQAN is slightly poorer than in Nelson et al. [34]. Besides, the imbalance of the preprocessed dataset may be another factor. In future researches, we will expand the size of the dataset and add more emotional information to the images of it to better train and evaluate the VQA models integrated with affective computing.

Table 2. The results of AVQAN and DAVQAN on our preprocessed dataset

Model	**Accuracy**
AVQAN	47.83%
DAVQAN	55.43%

4.5. AVQAN and DAVQAN with GloVe

Feature representation plays an important role in improving VQA performance. The AVQAN and DAVQAN described above use the one-hot representation of words to embed the question words. Now there are more advanced methods for word embedding, such as GloVe [28], ELMo [29] and BERT [7]. In order to study whether the more advanced word embedding methods can improve the performance of the two models, we use the GloVe word embeddings to replace the one-hot representation of the question words and carry out experimental verification. GloVe is a global log-bilinear regression model for the unsupervised learning of word representations, which can directly obtain the global corpus statistics. Instead of using individual context windows in a large corpus and the entire sparse matrix, the GloVe model uses the nonzero elements in a word-word co-occurrence matrix to train and construct a vector space with meaningful sub-structure thus efficiently leverages statistical information.

Specifically, instead of using the one-hot encoding, we use the 300-D GloVe word embeddings pre-trained on a large-scale corpus to transform the question words into 300-dimensional word vectors. The following operations are the same as those in the original AVQAN and DAVQAN models, that is, we embed the question word vectors and the image features into the same dimension and then take them as input to the LSTM model to complete the subsequent experiments. Table 3 shows the results of AVQAN and DAVQAN with GloVe. By comparing with Table 2, we can see that the accuracy of AVQAN and DAVQAN models after using GloVe is improved by 3.8% and 2.72% respectively, which proves that the improvement on the word embedding method can improve the performance of VQA models integrated with affective computing.

Table 3. The results of AVQAN and DAVQAN with GloVe on our preprocessed dataset

Model	**Accuracy**
AVQAN with GloVe	51.63%
DAVQAN with GloVe	58.15%

4.6. AVQAN and DAVQAN with ResNet

The AVQAN and DAVQAN described above use the VGGNet [36] to extract image features. Now there are more advanced image feature extractors, such as the ResNet [18] and the bottom-up attention network [2]. In this section, we use the ResNet to replace the VGGNet to extract more fine-grained image features to study whether the more advanced image feature extractors can improve the performance of the two models. The depth of image representation is crucial to many visual tasks, but the deeper neural networks are more difficult to train. The ResNet uses a residual learning framework to simplify the training of networks that are substantially deeper than those used previously. Instead of learning unreferenced functions, the ResNet explicitly reformulates the layers as learning residual functions with reference to the layer inputs. Empirical evidence shows that these residual networks are easier to optimize and can obtain accuracy from significantly increased depth to produce better results than previous networks.

The original AVQAN and DAVQAN models use VGGNet to extract image features to infuse attention mechanism and complete mood detection. For simple and convincing comparison, we only use the ResNet to replace the VGGNet used in the mood detector to complete our experiments. The rest of the two models are the same as the corresponding original model and the results are shown in Table 4. By comparing with Table 2, we can see that the accuracy of AVQAN and DAVQAN models after using ResNet-50 is improved by 5.97% and 1.09% respectively, which proves that better image feature extractors can improve the performance of VQA models integrated with affective computing. We have also explored the deeper network ResNet-101 and use it to improve the accuracy of AVQAN model to 54.35%. For DAVQAN, using ResNet-50 and ResNet-101 yielded similar results.

5. Conclusion and future work

The Affective Visual Question Answering Network (AVQAN) enriches the model's understanding and analysis of VQA by making use of the emotional information contained in the images while maintaining the same level of accuracy as ordinary VQA baseline

Table 4. The results of AVQAN and DAVQAN with ResNet-50 on our preprocessed dataset

Model	Accuracy
AVQAN with ResNet-50	53.80%
DAVQAN with ResNet-50	56.52%

models. It is a fairly new task to integrate the emotional information contained in the images into VQA. In AVQAN, however, it is difficult to separate question-guided-attention from mood-guided-attention due to the concatenation of the question words and the mood labels. We think that this kind of concatenation harms the performance of the model. To mitigate this effect, we propose the Double-Layer Affective Visual Question Answering Network (DAVQAN), which divides the task of generating emotional answers in VQA into two relatively simple subtasks, i.e., the generation of non-emotional answers and the generation of mood labels, and uses two independent layers to tackle the two subtasks respectively. Although the word embedding method, the feature extractors, and the attention mechanism used by the two models are the same, the overall performance of DAVQAN is 7.6% higher than that of AVQAN. We also introduce more advanced word embedding method and more fine-grained image feature extractor into AVQAN and DAVQAN to further improve their performance and obtain better results than their original models, which proves that VQA integrated with affective computing can improve the performance of the whole model by improving these two modules just like the general VQA. Furthermore, the performance of the models is limited because the dataset used is not large enough to give full play to the advantages of deep learning, and the emotional information contained in the images of the dataset is not rich enough. In future work, we will construct a larger, more specialized, and more balanced dataset to promote VQA tasks integrated with affective computing.

References

1. Agrawal, A., Jiasen, L., Antol, S., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D.: Vqa: Visual question answering. ICCV (2015)
2. Anderson, P., Xiaodong, H., Buehler, C., Teney, D., Johnson, M., Gould, S., Lei, Z.: Bottom-up and top-down attention for image captioning and visual question answering. CVPR (2018)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. IJCAI pp. 2670–2676 (2007)
4. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N.D., Blei, D.M., Jordan, M.I.: Matching words and pictures. The Journal of Machine Learning Research pp. 1107–1135 (2003)
5. Chen, K., Dahua, L., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? text-to-image coreference. CVPR pp. 3558–3565 (2014)
6. Chowdhury, M.I.H., Sridharan, S., Fookes, C., Nguyen, K.: Hierarchical relational attention for video question answering. ICIP (2018)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018)
8. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences for images. ECCV pp. 15–29 (2010)
9. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. EMNLP (2016)

10. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. ECCV pp. 241–257 (2016)
11. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. CVPR (2016)
12. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. ICML 37, 1462–1471 (2015)
13. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zeroshot recognition. ICCV pp. 2712–2719 (2013)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997)
15. Jia, D., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. CVPR pp. 785–792 (2011)
16. Jiang, Y., Liang, W., Tang, J., Zhou, H., Li, K.C., Gaudiot, J.L.: A novel data representation framework based on nonnegative manifold regularisation. Connection Science (forthcoming)
17. Junyoung, C., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS (2014)
18. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
19. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. The 3rd International Conference for Learning Representations (2015)
20. Krishna, R., Yuke, Z., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li-Jia, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. CoRR abs/1602.07332 (2016)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NIPS 1, 1097–1105 (2012)
22. Kulkarni, G., Premraj, V., Dhar, S., Siming, L., Yejin, C., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating simple image descriptions. CVPR pp. 1601–1608 (2011)
23. Li, K., Jiang, H., Zomaya, A.: Big data: Management and processing. Taylor & Francis (2017)
24. Li, K., Martino, B.D., Yang, L.T., Zhang, Q.: Smart data: State-of-the-art perspectives in computing and applications. Taylor & Francis (2019)
25. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. NIPS 1, 1682–1690 (2014)
26. Martino, B.D., Li, K., Yang, L., Esposito, A.: Internet of everything: Algorithms, methodologies, technologies and perspectives. Springer (2018)
27. Mitchell, M., Dodge, J., et al., A.G.: Midge: Generating image descriptions from computer vision detections. EACL pp. 747–756 (2012)
28. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. EMNLP pp. 1532–1543 (2014)
29. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. NAACL (2018)
30. Pirsiavash, H., Vondrick, C., Torralba, A.: Inferring the why in images. CoRR abs/1406.5472 (2014)
31. Quanzeng, Y., Hailin, J., Jiebo, L.: Visual sentiment analysis by attending on local image regions. AAAI (2017)
32. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people with ”their” names using coreference resolution. ECCV pp. 95–110 (2014)
33. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B.: Translating video content to natural language descriptions. ICCV pp. 433–440 (2013)
34. Ruwa, N., Qirong, M., Liangjun, W., Ming, D.: Affective visual question answering network. MIPR (2018)
35. Shaoqing, R., Kaiming, H., Girshick, R., Jian, S.: Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS 1, 91–99 (2015)

36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2014)
37. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng., A.Y.: Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics (2014)
38. Tsung-Yi, L., Maire, M., et al., S.B.: Microsoft coco: Common objects in context. ECCV (2014)
39. Voorhees, E.M., Tice, D.M.: Building a question answering test collection. SIGIR pp. 200–207 (2000)
40. Wang, Q., Zhu, G., Zhang, S., Li, K.C., Chen, X., Xu, H.: Extending emotional lexicon for improving the classification accuracy of chinese film reviews. Connection Science (forthcoming)
41. Xinlei, C., Hao, F., Tsung-Yi, L., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. CoRR abs/1504.00325 (2015)
42. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. ICML (2015)
43. Yuke, Z., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. CVPR (2016)
44. Zhang, S., Hu, Z., Zhu, G., Jin, M., Li, K.C.: Sentiment classification model for chinese microblog comments based on key sentences extraction. Soft Computing (forthcoming)
45. Zhou, Y., Jun, Y., Chenchao, X., Jianping, F., Dacheng, T.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Transactions on Neural Networks and Learning Systems pp. 5947–5959 (2018)
46. Zitnick, C.L., Parikh, D., Vanderwende, L.: Learning the visual interpretation of sentences. ICCV pp. 1681–1688 (2013)

Zihan Guo is currently pursuing the Ph.D. degree in Shanghai Maritime University. His research interests include computer vision and natural language processing methods related to visual question answering.

Dezhi Han received the Ph.D. degree from the Huazhong University of Science and Technology. He is currently a Professor of computer science and engineering with Shanghai Maritime University. His research interests include network security, cloud computing, mobile networking, wireless communication, and cloud security.

Kuan-Ching Li is a Professor in the Dept of Computer Science and Information Engineering (CSIE) at Providence University, Taiwan, where he also serves as the Director of the High-Performance Computing and Networking Center. He published more than 300 scientific papers and articles and is co-author or co-editor of more than 25 books published by Taylor & Francis, Springer, and McGraw-Hill. Professor Li is the Editor in Chief of the Connection Science and serves as an associate editor for several leading journals. Also, he has been actively involved in many major conferences and workshops in program/general/steering conference chairman positions and has organized numerous conferences related to computational science and engineering. He is a Fellow of IET and a senior member of the IEEE. His research interests include parallel and distributed computing, Big Data, and emerging technologies.

Received: May 15, 2020; Accepted: September 10, 2020.