# Human-centric Computing and Information Sciences

March 2021 Volume 11

www.hcisjournal.com

KIPS Korea Information Processing Society

KIPS CSWRG Korea Information Processing Society Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

**RESEARCH(Research Manuscript)** **Open Access**

Human-centric Computing and Information Sciences (2021) 11:11
DOI: https://doi.org/10.22967/HCIS.2021.11.011
Manuscript received June 28, 2020; accepted January 18, 2021; published March 15, 2021.

# Visual Question Answering Research on Multi-layer Attention Mechanism Based on Image Target Features

Danyang Cao[1,2,*], Xu Ren[1], Menggui Zhu[1], and Wei Song[3]

## Abstract

Visual question answering (VQA) aims to output a natural language answer based on a picture and a related question in order to achieve machine language understanding. Numerous approaches have been proposed to solve this problem, mainly inspired by methods of natural language processing and deep learning. Still, existing approaches always use the convolutional neural network (CNN) to extract the image features and employ the recurrent neural network to extract the features of the question sentences. The two methods are accompanied by the single-layer attention mechanism to improve the accuracy of the visual question-and-answer model, but these approaches perform poorly in the visual scene understanding and reasoning of knowledge about the image. In this study, we first used the Faster RCNN to extract the target features as the feature representation of the entire image. Second, we utilized the multi-layer attention mechanism to improve the model accuracy to deal with the challenges faced by current methods. Experiments proved that our suggested framework improves the accuracy of the task of visual question answering on the VQA V2 dataset, showing significant improvements in the trade-off between accuracy and speed.

## Keywords

Visual Question Answering, Multi-layer Attention Mechanism, Long Short-Term Memory, Convolutional Neural Network

# 1. Introduction

Deep learning has become one of the fastest growing and most exciting machine learning areas in the world. It has also brought huge profits to enterprises in industrial applications in recent years. As an emerging task connecting the two deep learning areas of computer vision and natural language processing [1], visual question answering (VQA) has been applied to help blind/visually impaired users understand visual information and to provide users with the required image information on the network or social media. The most important application is the integration of VQA systems into image retrieval systems, bringing huge profits to social media and e-commerce. VQA is a natural language question and answer about visual images, and its definition in the source paper can be summarized as follows: a VQA system takes as input an image and a free-form, open-ended, natural language question about the image and


**Corresponding Author:** Danyang Cao (ufocdy@163.com)
[1]School of Information Science and Technology, North China University of Technology, Beijing, China
[2]Beijing Key Laboratory on the Integration and Analysis of Large-scale Stream Data, Beijing, China
[3]International School, North China University of Technology, Beijing, China

produces a natural language answer as output [2]. A visual question answering system input is a picture and a free-form, open-ended natural language question about the picture. The output is a natural language answer. This definition is a good interpretation of what is VQA. In a nutshell, VQA is asking questions based on pictures.

Nowadays, the application of multimedia is becoming increasingly extensive. In order to meet the increasing demands of people, image question answering [3, 4], speech-based VQA [5], and video question answering [6, 7] technologies have appeared one after another. With the rapid development of mobile phones and networks, users need to take photos with their mobile phones to achieve online Q&A [8]. People have been trying to reduce the impact of ambiguity in natural language [9].

It is easier for people to read the pictures and question statements and then output the answers, but it is difficult for machines without thoughts. How to make a machine have a person's "thought" and understand the information contained in images and question sentences has become a problem that has been solved by visual question answering researchers in recent years [3]. This requires a variety of artificial intelligence techniques such as object recognition, target detection, fine-grained recognition, behavior recognition, and textual understanding in natural language processing. With the continuous exploration of researchers at home and abroad, visual question answering technology can be divided into several categories, and this will be covered in Section 2.

Our research focused on the deep learning model. According to our research, object recognition from an image is done by the convolutional neural network (CNN). The recurrent neural network (RNN) with long short-term memory (LSTM) cell has outstretched the bar on sequence prediction jobs as well as machine translation [10, 11]. The researchers directly combined both networks and trained end to end to generate the answer [12, 13]. Still, this kind of approach is not so good as it is only able to answer common and simple questions related to the image's content, i.e., "What is the color …?" or "How many?" To find a correct answer for the different types of questions, understanding of an image should be different [14, 15]. VQA needs various types of understanding of an image, not only caption generation or image entity recognition but visual scene understanding and reasoning of knowledge about the image. The approaches we discuss in Section 3 took the VQA problem as a classification task, in addition to some network architectures designed to solve the abovementioned problem in a joint embedding manner. To overcome the previous problem, we used the Faster region-based CNN (RCNN) for feature extraction of image. According to [16], the image target feature can replace the information well rather than the whole image. Thus, we proposed using the image target feature as image information representation.

According to recent research [17, 18], the relevant image features can be selected by using the attention mechanism in the neural network, which can improve the accuracy of the VQA model. Nonetheless, the single-layer attention mechanism has a limited effect on the improvement of the model. This paper uses the multi-layer attention mechanism to improve the problem of the single-layer attention mechanism being insufficient to solve the task of VQA.

Our approach makes several major contributions. First, we used Faster RCNN to extract the target features as the feature representation of the entire image. Second, we used a multi-layer attention mechanism to improve the accuracy of the model and end-to-end training. We have obtained higher scores on the VQA V2 dataset compared to other current research methods.

# 2. Related Research

With the continuous exploration of researchers at home and abroad, VQA technology can be divided into the following categories.

## 2.1 Multinode System

### 2.1.1 Prediction method for the answer type

In 2016, Kafle and Kanan [19] proposed a Bayesian model of VQA. The idea of this method is to model the statistical features of image features and questions together into a way of inferring the relationship between questions and images. The authors used the features of the questions and the types of answers to model the probability of image features. They also introduced several simple baseline methods, such as send only question features or answer features to a logistic regression model, or send both features to logical regression, etc. Their work was evaluated on a VQA data set.

### 2.1.2 Multi-world question answering

The model is analyzed from the question statement, and the semantic analysis tree is obtained. The additional features are then obtained from the original image or the image segmentation block. Finally, the deterministic evaluation function is used to evaluate the probability function, and then the simple logarithmic linear model is used to obtain the probability of hiding a variable based on questions. In view of the uncertainty of the model's segmentation and classification labels, the authors further extended the model to a multi-world world [20].

## 2.2 Deep Learning Method

Most of the VQA models based on deep learning methods use the CNN [21, 22] to process images and obtain image features. LSTM [23, 24], bi-directional LSTM (BLSTM) [18, 25], and gated recurrent unit (GRU) [15, 26, 27] are used to process the question statement to obtain the question feature, and then combine the image feature and the question feature in different ways and obtain the answer after processing. The application of attention mechanism has been very successful in machine translation. It has been used by many researchers in image caption and VQA. The attention mechanism can focus on the important part of the image or question, so that the model pays more attention to these parts when extracting features, which is beneficial to the improvement of experimental accuracy. Shih et al. [28] proposed an attention-based model called WTL (where to look). The authors used the VGG network to encode the image. The question feature is averaged by the word vector in the question. Determination as to which position in the image is more important is done by calculating the attention vector on the image features. Finally, the attention vector is weighted onto the image feature and connected to the question, and then sent to the dense + softmax layer to get the answer.

## 2.3 Other Models

There are some other models, such as the neural network block model proposed by Andreas et al. [29] and the AMA (ask me anything) model proposed by Wu et al. [30]. Both use more ideas and combine more techniques, not just the attention of image features or question features.

Deep learning technology has achieved widespread success in computer vision and natural language processing. With its powerful feature learning ability, it eliminates the need to select features manually; thus greatly reducing the workload of manual operations. The use of CNNs to extract image features and RNN processing text data has achieved great success in the field of image annotation. Unlike image annotation tasks that generate only a descriptive sentence for the image, the VQA task needs to read not only the picture content but also the question statement and, according to the question statement, combine the information in the image to get the answer. This task is more difficult, with higher technical requirements. At present, the better model is still combining the CNN and RNN. Most of the researchers also improve on this model.

Our research focused on the deep learning model. Antol et al. [1] proposed a model called Deeper

LSTM Q + norm I, which first uses depth LSTM to extract the features of the question statement according to the word in the sentence, followed by the CNN to extract the image features and use them to represent the whole image information and subsequently regularize the image features to improve the feature quality. After that, the image features and the question statement features are merged into the same vector space. Finally, the fusion features are sent to the multi-layer perceptron to generate the answer. This is the earliest idea on using the deep learning method to deal with the VQA task. Although the model is not so good due to the single model structure, it provides a good foundation for the later researchers to study the VQA task.

Inspired by the method proposed in [1], Ren et al. [31] proposed another structure that uses CNNs to extract image features and utilizes image features as a word in the question statement (may be the first word or the last word) and sends these words to LSTM for the same processing. This method uses RNN and visual semantic embedding in the middle phases of image segmentation and object detection. In this paper, their main contribution is the question generation algorithm that transforms the image descriptive dataset into question-answer format. They treated the VQA problem as classification rather than answer generation. Still, this approach is only able to answer common and simple questions related to the image's content, i.e., "What is the color …?" or "How many?" Cho et al. [27] improved the model proposed in [31] and proposed a model similar to sequence-to-sequence (Seq2Seq). They still used convolutional neural networks to extract image features. Unlike the paper in [31], the authors sent the image features together and the question word vector to the LSTM and used the LSTM network to generate the final answer. This method realizes the Seq2Seq structure. Although the image features and question features are sent to the LSTM at each moment, it seems to improve the utilization of the image features. In fact, taking the image features as input does not determine whether they work. In this paper, their main contribution is primarily using the Seq2Seq model to solve the problem of machine translation. They proposed a novel neural network model called RNN encoder–decoder that consists of two RNNs. One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. Nonetheless, this approach lacks various types of understanding of an image, not only caption generation or image entity recognition but visual scene understanding and reasoning of knowledge about the image.

Since the attention mechanism has been successfully applied to machine translation and image annotation generation tasks [12, 32], Shih et al. [28], Zhu et al. [33], and Lu et al. [34] proposed attention mechanisms of different structures and successfully improved the accuracy of the VQA model. Attention mechanism can increase the computational efficiency of the VQA model by introducing attention mechanism to increase the weight of specific parts (images and/or questions) in the input vector. The main idea of the attention mechanism is to replace the overall (image range) features with spatial feature maps and to allow interaction between the question and specific areas of these maps.

According to the previous work, VQA can be split into two branches: image question answering (image QA) and video question answering (video QA), both of which have attracted considerable attention in the past few years. In these two sub-domains, a considerable number of datasets have been crafted and published to benefit the community, either through crowdsourcing manner on Amazon Mechanical Turk [35] or weakly supervised algorithms [36, 37].

An intelligent agent designed for VQA is required to understand fully the textual information in the question and visual information in the image or video and leverage an effective fusion mechanism to find the semantic connection in these two modalities so as to return the correct answer. An end-to-end framework normally consists of three stages: feature extraction, feature fusion, and answer generation.

In this study, we used target area features instead of the whole image. The characteristics of the target area in the image are proven to be high-quality image features that can improve the classification or detection accuracy of the model and promote the model effect of the image annotation task. This study used high-quality image features to replace the whole image features used in the previous work, which can eliminate the problem of insufficient image feature quality caused by poor image quality. Moreover, for better quality of the image features, we regularized the image features. For the processing of question statements, we still used the LSTM network, which works well in text processing. In connecting image features and question features, this study used a multi-layer attention mechanism. Such multi-layer attention mechanism focuses on more specific targets in the image, improving the sub-optimal result of the single-layer attention mechanism when the target is too large. The multi-layer attention mechanism further reduces the scope of attention compared with the single-layer attention mechanism, so that the image region features are fully utilized in the model training. Furthermore, the degree of correlation between the image and the question statement is improved, which can help us generate more accurate answers and improve model performance.

# 3. Proposed VQA Models

This study implemented a VQA task based on the deep learning method. Our model structure is shown in Fig. 1. In this study, the feature of the target region in the image was used instead of the global feature of the image, which improves the image feature quality and performance of the model. A method of word embedding is used for question features in order to feed it better into the LSTM network. Most importantly, we proposed improvements in the use of attention mechanisms that use a multi-layer attention mechanism to focus more specifically on a certain part of the image's features, resulting in a better model.

In the next section, we will describe our approach based on the processing of image features, processing of question text data, multi-layer attention mechanism, and generation of the final answer.

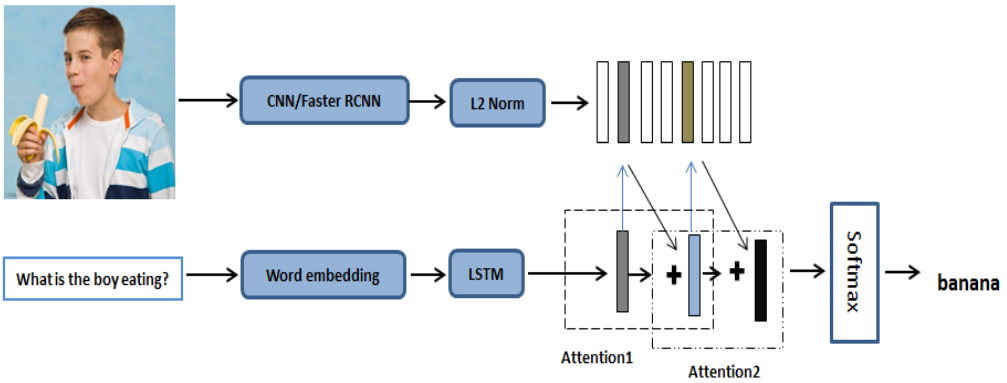


**Fig. 1.** Overall structure of the model.

## 3.1 Image Feature Processing

The image is extracted by the CNN. Commonly used image feature extraction networks include VGG series, Inception series, ResNet series, etc. In order to obtain high-quality image features, we used the ResNet network as a feature extraction network. The features extracted here are represented as a matrix

of $K \times 2048$; each vector size is 2048 dimensions, each image has $K$ vector representations, and $K$ represents the different positions in the image as shown in Fig. 2.

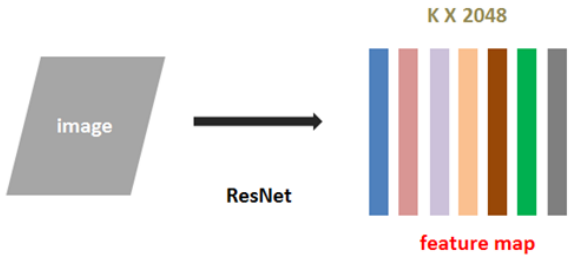


**Fig. 2.** Feature map of the convolutional neural network extraction.

Referring to the method proposed by Anderson et al. [14], bottom-up attention is used to obtain image features. This method uses the ResNet network to extract features of the image, and then employs the Faster RCNN framework to select the image target position. The final result is a feature map of top-K targets mapped by features generated by ResNet. For convenience of subsequent work, we fixed $K$ to 36, which means that each target image selects 36 target areas as the final image features and each target is a 2048-dimensional vector.

In order to train the model more conveniently, and for greater convenience of use in combination with the question features, we first initialized the image features in order to convert the dimensions of the image features into the same size as the question vector.

$$v_I = tanh(W_I f_I + b_I) \tag{1}$$

$v_I$ is a matrix, each column of which is a visual feature vector of target area $i$; $f_I$ is an image feature representation of each picture, and $W$ and $b$ are related parameters.

## 3.2 Question Processing Model

For the VQA model, the processing of the question statement is very important because it directly affects the training effect of the question model. The preprocessing of the vocabulary in the answer is equally important, since it determines what kind of answers your model will generate.

The use of deep learning methods to deal with question statements in visual question answering tasks can achieve good results; in fact, the use of RNNs to process text data has become a more mature technology in the field of natural language processing. In order to understand better the characteristics of the question statement and make the question statement feature play a better role in generating the answer, this study used the LSTM network to process question statements. LSTM can remember the long-term information, mainly solving the problem of long-term dependence in RNN.

The part of the LSTM structure that remembers long-term information is called the memory unit and is represented by $c_t$, which is the most critical part of the entire structure because it is similar to a conveyor belt and it rarely interacts linearly with other parts of the network. This makes saving information easy. LSTM contains forgetting gates, input gates, and output gates. It is these special door structures that determine how information is selected and passed. In each time step, LSTM will accept input vector $x_t$, which is the word vector, and then update the value of memory unit $c_t$ and output hidden layer state $h_t$. In LSTM, the information update process is controlled by the gate mechanism. Forgetting gate f controls how much information in previous time $c_{t-1}$ will be retained; input gate $i_t$ controls how much information in current input $x_t$ is updated to the memory unit. Output gate $o_t$ controls how much information in the memory unit is sent to the hidden layer state for subsequent output. The information update process in LSTM is as follows:

$$i_t = \sigma(W_{x_i}x_t + W_{h_i}h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_{x_f}x_t + W_{h_f}h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(W_{x_o}x_t + W_{h_o}h_{t-1} + b_o) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t \ tanh(W_{x_c}x_t + W_{h_c}h_{t-1} + b_c) \quad (5)$$

$$h_t = o_t \ tanh(c_t) \quad (6)$$

Here, *i*, *f*, *o*, and *c* represent the input gate, forgetting gate, output gate, and memory unit. Weight matrix $W$ and bias term $b$ are parameters learned by LSTM in training. Better training is beneficial to the correction of parameters, improving the accuracy of the model.

Given a question statement $q = [q_1, q_2, \ldots\ldots\ q_T]$, $q_T$ represents the one-hot code of the word. In this study, an embedded matrix was used to embed the word into a vector space, and the embedding process can be expressed as $x_t = W_e q_t$. At each time step, we can send the words one by one into the LSTM structure:

$$x\_t = W\_e \ q\_t \ , t \in \{1, 2, \ldots\ldots T\} \quad (7)$$

$$h_t = LSTM(x_t), \ \ t \in \{1,2, \ldots\ldots T\} \quad (8)$$

As shown in Fig. 3, the question statement "What is the boy eating" is being sent to LSTM, and the last hidden layer vector is used as the representation vector for this question. LSTM can represent the question as vector $v_Q$, which can be used for later research.

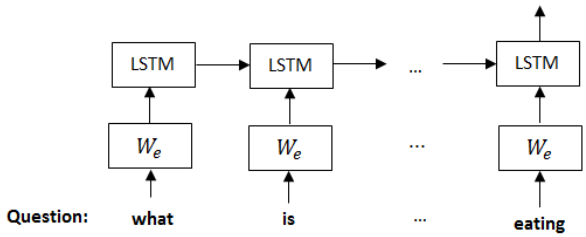


**Fig. 3.** Question statement processing model.

## 3.2 Multi-layer Attention Mechanism

Attention mechanism has played an important role in the field of natural language processing such as machine translation, greatly improving the robustness of related models in this field. In recent years, it has been used in the field of image annotation, and it has also yielded very good results. In the image annotation task, the use of the attention mechanism makes the model pay more attention to the relevant area of the image when generating the statement, which is a kind of mapping of human observation things. Applying the attention mechanism to the field of visual question answering will also enhance the effect of the model.

The feature representation matrix of the image can be obtained from the two sections above, which is called $v_I$; the feature representation vector of the question can also be obtained, which is called $v_Q$. The attention mechanism links the image feature representation with the question feature representation. In order to narrow down further the range of image-related regions of interest to the model, this study used a multi-layer attention mechanism that adds a layer of attention mechanism based on the first layer of attention mechanism to achieve the goal of determining the target range.

In many cases, the single-layer attention mechanism is not able to determine the area of interest for the image. For example, in Fig. 4(a), the boy is eating bananas, and the single-layer attention mechanism determines the extent of the white blurred area in the graph, but this enlarges the actual range of the model. If there are many targets in the image, it may bring wrong results. In order to improve the accuracy of the model, this study used a multi-layer attention mechanism based on the single-layer attention mechanism. The region of interest can be accurately located as shown in Fig. 4(b) so the model gradually eliminates noise, and the final located region is highly correlated with the generated answer. The model is more accurate.

(a) (b)

**Fig. 4.** (a) Single-layer attention mechanism and (b) multi-layer attention mechanism.

After determining image feature representation $v_I$ as well as the question statement region represented by the symbol $\widetilde{v_i}$, the summation formula is shown in Equation (9). Then we need to combine $\widetilde{v_i}$ with problem vector $v_Q$ to form a query vector, denoted by $u$. The process is shown in Equation (10). Vector u can be regarded as a refined query vector because it encodes the visual information and the problem information. This code is closely related to the underlying answer.

For feature representation $v_Q$, you can use a single-layer neural network and a softmax function to generate the attention distribution of the first image region, which is the first layer of the attention mechanism. The relevant formula is as follows:

$$h_A = tanh\left(W_{I,A}v_I \oplus (W_{Q,A}v_Q + b_A)\right) \tag{9}$$

$$p_I = softmax(W_P h_A + b_P) \tag{10}$$

In the formula $v_I \in R^{d \times m}$, $d$ is the dimension of the image feature representation, $m$ is the number of regions in each image, and $v_Q \in R^d$ is a $d$-dimensional vector. Here, $W_{I,A}, W_{Q,A} \in R^{k \times d}$, $W_P \in R^{1 \times k}$, and $P_I \in R^m$ are the m-dimensional vector corresponding to the attention probability of each region in $v_Q$. We add a vector to the matrix with the symbol $\oplus$. Since $W_{I,A}v_I \in R^{k \times m}$, and $W_{Q,A}v_Q$ and $b_A$ are both vectors, the addition between the matrix and the vector is needed to operate each column of the matrix.

The attention distribution of the image region is obtained. It is necessary to calculate the weighted sum of the image vectors of each.

$$\widetilde{v_I} = \sum_i p_i v_i \tag{11}$$

$$u = \widetilde{v_I} + v_Q \tag{12}$$

Compared with the simple combination of problem vectors and image global features, the advantage of the attention mechanism is that the visual regions that are more relevant to the problem are given

higher weights, so u contains more information, which is beneficial for the model to generate a more accurate answer. Nonetheless, the single-layer attention mechanism may not be sufficient to locate the correct region in the image. Therefore, this study used a multi-layer attention mechanism, which is actually the result of iteration. Each layer of the attention mechanism extracts finer-grained visual information for the generation of the answer; the formula of the multi-layer attention mechanism is as follows, where *k* represents the *k*-th layer attention mechanism:

$$h_A^k = tanh\left( W_{I,A}^k v_I \oplus \left( W_{Q,A}^k u^{k-1} + b_A^k \right) \right) \tag{13}$$

$$p_I^k = softmax\left( W_P^k h_A^k + b_P^k \right) \tag{14}$$

Here, $u^0$ is initialized by $v_Q$, with the aggregated image feature vector then added to the previous query vector to form a new query vector; thus completing the update of the query vector. The update formula is as follows:

$$\tilde{v}_I^K = \sum_i p_i^k v_i \tag{15}$$

$$u^k = \tilde{v}_I^K + u^{k-1} \tag{16}$$

As can be seen from the formula above, at each attention mechanism layer, we used the joint question and image feature vector $u^{k-1}$ as the query vector to query the image, and then updated the new query after selecting the image region. The vector is updated by the formula $u^k = \tilde{v}_I^K + u^{k-1}$. In theory, this process can be updated *K* times, and then final vector $u^k$ is used to infer the final answer. The inference formula is as follows:

$$P_{answer} = softmax(W_u u^K + b_u) \tag{17}$$

In Fig. 4(a) and 4(b), it can be seen that, after the first layer of attention mechanism processing, the model can roughly infer the area that needs to be selected, but it is not accurate. After the second layer of attention mechanism processing, the model is more clearly focused on the area corresponding to the answer to get the correct answer.

# 4. Experiments

## 4.1 Dataset

This study used the VQA V2 dataset commonly used in visual question answering. It is manually labeled and divided into image data and text data. The image data is mainly composed of two parts: real image and cartoon image. In this study, the real image of the COCO image annotation data set was used as experimental data. The training data has 82,783 images, the verification set has 40,504 images, and the test set has 81,434 images. There are three questions for each image, and each question has 10 answers marked by a manual annotator. There are 248,349 training questions and 121,512 verification questions in the data set. In this study, we used the top 1,000 answers that occur the most as an output set. These answers account for 82.67% of all answers, and they are basically able to answer common questions. We used a validation set for local testing, and the results are shown in Section 4.3.

## 4.2 Model and Training Parameter Configuration

For the feature representation of each image, this study used a size of 36 × 2048, representing 36 targets per image; each target is a 2048-dimensional vector, which can be more abundantly represented by high-dimensional vectors. For the question model, if the length is not enough, it is filled with 0. If the length

is greater than 26, only the first 26 words are intercepted. Basically, the problem sentences are less than 26 words long. Each word is represented by a 512-dimensional vector for better representation of vocabulary information and can be better sent to LSTM for calculation.

For the model structure, two layers of LSTM are used for calculation. The size of each layer of LSTM structure is 256, and the size of the joint embedding vector (embedding image features and problem features into the joint space) is 1024. In training, the learning rate is first set to 0.0003, and the learning rate attenuation value is set to 0.999975, which is based on training experience. The training batch size is 100, i.e., each training will train different 100 samples, which can better prevent the model from overfitting. At each training session, the probability of dropout per neuron connection is 0.5, which means that half of the neurons in training are not connected, and the connection of neurons changes every time you train. It is also an effective means of preventing model overfitting.

## 4.3 Results and Discussion

In the VQA task are accuracy and WUPS (Wu-Palmer similarity) for the evaluation of the model. Accuracy seems to be a very straightforward evaluation criterion, and it can give correct scores on certain identified issues; for some open answers, however, accuracy cannot be used to make a correct judgment. For example, the answer to a question is “oak.” If the model gives the answer to the “tree,” then this answer cannot be said to be absolutely wrong; if the question is “what animal is in the picture?” and there are dogs, cats, and rabbits in the picture, then it is also an ambiguous question as to whether the answer “dog and cat” is correct. Thus, in order to solve these problems and evaluate different methods as accurately as possible, this type of problem must be solved. WUPS can solve this kind of problem to the greatest extent.

The WUPS measurement standard was proposed by Malinowski and Fritz [20] in 2014 based on the WUP proposed by Wu and Palmer [38] in 1994. It estimates the semantic distance between the model output answer and the correct label. This is a value between 0 and 1. It relies on WordNet to calculate the similarity by using the distance in the semantic tree containing the model’s output answer and the correct label. In this way, for a single entry, the study can get the result.

In order to train better, you need high computer configuration. This study used a Titan XP graphics card with memory size of 12 GB. Using GPU training can make the model training faster. The hard disk size is 500 GB, and the memory size is 64 GB. A good hardware environment can make research better.

$$\mathrm{WUPS}(\text{oak tree}, \text{tree}) = 0.94$$
$$\mathrm{WUPS}(\text{dogs, cats and rabbits}, \text{cats and rabbits}) = 0.8$$

As with almost all semantic metrics, WUPS assigns more important values to completely unrelated terms; to solve this problem, the scale of a score below 0.9 was reduced by 0.1 times in this study. In many cases, WUPS is undeniably more suitable as a visual question answering evaluation standard than the classic accuracy evaluation indicators. Nonetheless, since it relies on semantic similarity, if the real answer is “black” or “green” or other colors, the model output answer “red” will get a very high score. Therefore, this study used the accuracy and WUPS indicators to evaluate the model.

In order to verify that the image characteristics of the target area proposed in this paper have improved the model effect, this study carried out a verification experiment. Under the same conditions, the global image features and the target area image features are used for comparison, and the model is then evaluated; the scores obtained on the evaluation indicators are shown in Table 1.

**Table 1.** Comparison of different feature effects of the model

| Method | Accuracy | WUPS0.9 |
|---|---|---|
| All-image feature | 58.4 | 66.3 |

| | | |
|---|---|---|
| Object image feature | **60.8** | **70.5** |

It can be seen from Table 1 that, in case other configurations are unchanged (the model parameters are unchanged, the training steps are unchanged, and the question statement feature extraction methods are unchanged), only the representation of the image features is changed; method 1 uses the traditional global image representation method, and method 2 utilizes high-quality image target feature representation. In the test data set, accuracy and WUPS evaluation indicator are used. It can be seen that, in the evaluation of accuracy, the target area image features of this study are better than the global image features. In the WUPS0.9 evaluation indicator, the method in this study is much better than the traditional method. Therefore, replacing the global image features with image target region features can improve the performance of the visual question answering model.

Similarly, in this study, a multi-layer attention mechanism was used to improve the performance of the model. In order to verify the improvement of the attention mechanism for the VQA model, we also carried out related experiments and modified the model to verify the performance of the non-attention mechanism, single-layer attention mechanism, two-layer attention mechanism, three-layer attention mechanism, and four-layer attention mechanism. Their performance on the evaluation indicators is shown in Table 2.

**Table 2.** Comparison of layers of different attention mechanisms

| Method | Accuracy | WUPS0.9 |
|---|---|---|
| No attention | 58.8 | 67.0 |
| One-attention | 59.3 | 68.2 |
| Two-attention | **60.8** | **70.5** |
| Three-attention | 60.0 | 69.9 |
| Four-attention | 60.8 | 70.3 |

As can be seen from Table 2, the attention mechanism can indeed bring some improvement to the model. The use of the attention mechanism is found to be much higher than the non-attention mechanism on the evaluation indicator, and the multi-layer attention mechanism performs better than the single-layer attention mechanism; thus, it is effective to use the multi-layer attention mechanism in the visual question answering task. As the number of attention mechanism layers increases, however, the model performance does not improve greatly. When the attention mechanism has three or four layers, the model training time becomes longer, and too many parameters are calculated, which leads to computer overloading for a long time, yet the model effect is not significantly improved. Therefore, after experimental research, the double-layer attention mechanism was adopted in this study. It can improve the performance of the model on the evaluation indicators.

The improved method proposed in this study is effective. In order to show that the research can achieve higher results in the VQA task, this study compared the method with the baseline method and other commonly used methods. The performance on the test data is as follows (Table 3).

**Table 3.** Comparison of different methods

| Method | Accuracy | WUPS0.9 |
|---|---|---|
| LSTM Q+I | 53.1 | 55.0 |
| VIS+LSTM | 55.7 | 65.5 |
| 2-VIS+LSTM | 58.9 | 66.8 |
| Our method | **60.8** | **70.5** |

Question: What's the woman eating?
Answer: pizza

Question: How many people?
Answer: 2

Question: What kind of sport is the woman playing?
Answer: tennis

Question: What color is the dog?
Answer: white

Question: Is this a creamy soup?
Answer: Yes

Question: What is to the right of the soup?
Answer: A
A. chopsticks and spoon B.Chopsticks
C. spoon D. shrimp

Question: What is the man doing in the street?
Answer: C
A.Walking B.Crossing
C.Crossing road D. Seeing

Question: Why is there a gap between the roof and wall?
Answer: D Correct answer:B
A.Ventilation B.airflow
C.provide air D.keep cow safe

**Fig. 5.** Model generation answer display.

As can be seen from Table 3, LSTM Q+I as a baseline method combines the global features of the image with the problem features and finally classifies them. The scores of the method on the evaluation indicators are 53.1 and 55.0, respectively, and the performance is not good. In the proposed VIS+LSTM and 2-VIS+LSTM methods, the performance of the model improved. Compared with these methods, however, the methods in this study have different degrees of improvement in accuracy and WUPS evaluation indicators. This study can be verified to have improved the performance of the VQA model.

This study used several pictures to verify the effect of the model, and the effect picture is shown above. As can be seen from Fig. 5, the model can achieve good results based on various question types, including open-ended problem, counting problem, Yes/No problem, and multiple-choice problem. Most of them

can get the correct answer by using our model; although there may be some deviations with the open-ended problem, this issue is not serious. In general, the algorithm can be applied to real life as seen from the experimental result. Although the accuracy of the model has improved, it still does not reach the basic level of human needs. There is also a need to continue improving the performance of the model in the follow-up work, and this is also the direction that researchers need to work on.

### 4.4 Summary

Through experimental verification, the target detection algorithm is used to extract the image features of the target area instead of the global image features in the visual question answering task. It improves the performance of the model on the evaluation indicators. At the same time, the use of multi-layer attention mechanism also improves the score of the model on the evaluation indicator to a certain extent; if the attention mechanism exceeds three layers, however, it cannot bring more obvious improvement to the model, and it also increases the number of parameters and the computational burden. Therefore, using the two-layer attention mechanism can not only improve the model effect; the model can also be trained normally without overloading the computer. In summary, this study improved the score of the VQA model in the evaluation indicator. The research is meaningful, but the VQA model can certainly achieve better results with the development of computer hardware and deep learning technology. Therefore, it can be better applied to the actual situation, bringing more value

## 5. Conclusion

In this study, we proposed a new method. First, we used the Faster RCNN to extract the target features as the feature representation of the entire image. Second, we employed a two-layer attention mechanism to improve the accuracy of the model. Experiments proved that this study improved the score of the VQA model in the evaluation indicator.

### Author's Contributions

Conceptualization, Song W. Investigation and methodology, Zhu M. Formal analysis, Ren X. Writing—original draft, review, editing, Cao D, Song W. All authors have read and approved the final manuscript.

### Funding

This study was funded by the Yuyou Talent Support Plan of North China University of Technology (No. 107051360019XN132/017), Fundamental Research Funds for Beijing Universities (No. 110052971803/037), Special Research Foundation of North China University of Technology (No. PXM2017_014212_000014), and Beijing Natural Science Foundation (No. 4162022).

### Competing Interests

The authors declare that they have no competing interests.

## References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 2425-2433.

[2] D. Teney, Q. Wu, and A. van den Hengel, "Visual question answering: a tutorial," *IEEE Signal Processing*

*Magazine*, vol. 34, no. 6, pp. 63-75, 2017.
[3] Y. Zhou, R. Ji, J. Su, Y. Wu, and Y. Wu, "More than an answer: neural pivot network for visual question answering," in *Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View, CA, 2017, pp. 681-689.
[4] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1367-1381, 2018.
[5] T. Zhang, D. Dai, T. Tuytelaars, M. F. Moens, and L. van Gool, "Speech-based visual question answering," 2017 [Online]. Available: https://arxiv.org/abs/1705.00464.
[6] H. Xue, Z. Zhao, and D. Cai, "Unifying the video and question attentions for open-ended video question answering," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5656-5666, 2017.
[7] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View, CA, 2017, pp. 1645-1653.
[8] W. Zhang, L. Pang, and C. W. Ngo, "Snap-and-ask: answering multimodal question by naming visual instance," in *Proceedings of the 20th ACM International Conference on Multimedia*, Nara, Japan, 2012, pp. 609-618.
[9] R. Li and J. Jia, "Visual question answering with question representation update (QRU)," *Advances in Neural Information Processing Systems*, vol. 29, pp. 4662-4670, 2016.
[10] P. Wang, Q. Wu, C. Shen, and A. van den Hengel, "The VQA-machine: learning how to use existing vision algorithms to answer new questions," 2016 [Online]. Available: https://arxiv.org/abs/1612.05386.
[11] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," 2016 [Online]. Available: https://arxiv.org/abs/1611.01646.
[12] 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 32nd International Conference on Machine Learning*, Lille, France, 2015, pp. 2048-2057.
[13] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual image question answering," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2296-2304, 2015.
[14] 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," 2017 [Online]. Available: https://arxiv.org/abs/1707.07998.
[15] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 30-38.
[16] N. Ruwa, Q. Mao, L. Wang, J. Gou, and M. Dong, "Mood-aware visual question answering," *Neurocomputing*, vol. 330, pp. 305-316, 2019.
[17] H. Xu and K. Saenko, "Dual attention network for visual question answering," in *Proceedings of ECCV 2016 2nd Workshop on Storytelling with Images and Videos (VisStory)*, Amsterdam, The Netherlands, 2016.
[18] Y. Lin, Z. Pang, D. Wang, and Y. Zhuang, "Task-driven visual saliency and attention-based visual question answering," 2017 [Online]. Available: https://arxiv.org/abs/1702.06700.
[19] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 4976-4984.
[20] M. Malinowsk and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1682-1690, 2014.
[21] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: a neural-based approach to answering questions about images," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1-9.
[22] J. Koushik, H. Hayashi, and D. S. Sachan, "Compositional reasoning for visual question answering," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017.
[23] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "FVQA: fact-based visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2413-2427, 2018.
[24] A. Jiang, F. Wang, F. Porikli, and Y. Li, "Compositional memory for visual question answering," 2015

[Online]. Available: https://arxiv.org/abs/1511.05676.

[25] A. Prakash, “Highway networks for visual question answering,” 2016 [Online]. Available: https://visualqa.org/static/slides/Highway%20Networks%20for%20VQA.pdf.

[26] H. Li, M. R. Min, Y. Ge, and A. Kadav, “A context-aware attention network for interactive question answering,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, 2017, pp. 927-935.

[27] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1724-1734.

[28] K. J. Shih, S. Singh, and D. Hoiem, “Where to look: focus regions for visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 4613-4621.

[29] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” 2015 [Online]. Available: https://arxiv.org/abs/1511.02799.

[30] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, “Ask me anything: free-form visual question answering based on knowledge from external sources,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 4622-4630.

[31] M. Ren, R. Kiros, and R. Zemel, “Exploring models and data for image question answering,” *Advances in Neural Information Processing Systems*, vol. 28, pp. 2953-2961, 2015.

[32] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.

[33] Y. Zhu, O. Groth, M. Bernstein, and F. F. Li, “Visual7w: grounded question answering in images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 4995-5004.

[34] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” 2017 [Online]. Available: https://arxiv.org/abs/1606.00061.

[35] K. H. Zeng, T. H. Chen, C. Y. Chuang, Y. H. Liao, J. C. Niebles, and M. Sun, “Leveraging video descriptions to learn video question answering,” in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 4334–4340.

[36] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and Y. Zhuang, “Video question answering via attribute-augmented attention network learning,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tokyo, Japan, 2017, pp. 829-832.

[37] M. Heilman and N. A. Smith, “Good question! Statistical ranking for question generation,” in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Los Angeles, CA, 2010, pp. 609-617.

[38] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 1994, pp. 133-138.