

A New Design of VQA System based on Weighted Contextual Features

Anonymous Submission

Abstract—Visual question answering (VQA) is a challenging task that requires a deep understanding of language and images. Currently, most VQA algorithms focus on finding the correlations between basic question embeddings and image features by using an element-wise product or bilinear pooling between these two vectors. Some algorithms also use attention models to extract features. In this paper, deeper analyses of these attention features are enabled by capturing their importance by weighting their contextual information. A novel interpretable VQA system leveraging weighted attention contextual features (WACF) is proposed for VQA tasks. This is a multimodal system which can assign adaptive weights to the features of questions and images themselves and to their contextual features based on their importance. Our new model yields state-of-the-art results on the MS COCO VQA datasets for open-ended question tasks.

I. INTRODUCTION

Visual question answering (VQA) has become a popular research problem that is being studied from the perspectives of multiple disciplines, such as natural language processing (NLP) and computer vision. The objective of a VQA task is to generate an answer based on an image and a given related question. The answer can be a number, a response of yes or no, or a word phrase. Such a task is not trivial because it is necessary to first understand each image and its corresponding question and then find the correlations between question-image pairs based on their own features as well as certain auxiliary external features [1], [2], [3], [4], [5], [6], [7], [8], [9]. Currently, most VQA approaches take advantage of the concept of multiple modalities by representing images and queries separately using two embedding or feature vectors [10]. In the vision mode, image features are extracted using a convolutional neural network (CNN) [11], whereas in the question understanding mode, a question embedding vector is generated to represent the semantic meaning of the question by using either a recurrent neural network (RNN) or the bag-of-words approach [12]. To identify the important information with regard to a question-image pair, most current algorithms use the concept of attention/co-attention. By assigning different weights to different image features, a good attention algorithm is able to select the features in an image that are most important in relation to the question being asked. There are many ways to generate attention weights, such as an elementwise sum or product or multimodal compact bilinear



Original Image Attention Map
Q: Is the boy wearing protective headgear
playing basketball?
Ans: Yes (Incorrect)

Fig. 1: An example in which an incorrect answer is obtained using a VQA model

pooling (MCB) [10], and most of them demonstrate reasonable performance on VQA tasks [13].

Despite the decent results that have been obtained by using different attention mechanisms for VQA, the overall performance achieved is still not comparable to that of human beings [6]. One possible reason for this shortfall is that humans can use more contextual information in both the question and the image to infer the answer. By observing the attention maps and test results generated by one current state-of-the-art VQA model, we can identify that most incorrect answers are generated for one of two reasons: 1. The question and its contextual information are not fully understood by the system; hence, the correct answer cannot be generated even though the correct regions of the image can be located.

2. Due to the attention mechanism applied, part of the important contextual information in the image is missing, adversely affecting the system performance.

In the example shown in Figure 1, the orange-colored text is contextual information that should be de-emphasized by the model since the text in purple is the real question that needs to be addressed. Consequently, although the correct regions of the image are emphasized in the attention map, the model still cannot find the correct answer.

To overcome these two potential difficulties, in this paper, a new interpretable multimodal structure for VQA is designed

by considering the contextual information in questions and images. A weighted contextual feature (WCF) structure is also proposed to balance the essential information from the question/image and the contextual information by assigning appropriate ratios through learning.

In this paper, our three main contributions are as follows:

1. We propose a framework in which contextual features extracted from multiple sources (image and query) are used to improve VQA performance by further considering the cross-impact of these features with different types of data.
2. We introduce interpretable multimodal WCF modules to balance the weights of different contextual features based on their importance.
3. Finally, we show that our model achieves state-of-the-art results on two well-known public VQA datasets.

The paper is organized as follows. Section 2 presents background on related work, including attention models, semantic contextual feature generation using long short-term memory (LSTM), and image contextual feature generation using multi-dimensional LSTM. A detailed system design based on contextual information and the proposed multimodal attention-based WCF model are introduced in section 3. Section 4 reports several experiments performed to demonstrate the performance of our new system in comparison to previous state-of-the-art results on two VQA datasets.

II. RELATED WORK

A. Attention-based Encoder-Decoder Model

The attention concept has been widely applied for model building in many domains, such as neural machine translation, the slot-filling task and sentiment classification in NLP [14], [15], [16] as well as the object detection and image captioning tasks in computer vision [17], [18]. As in [15], an attention mechanism can be used in an encoder-decoder structure to generate semantic contextual features. A sentence is passed through an RNN encoder word by word to generate the corresponding hidden states h_i . The contextual information is then captured by a weighted sum of the RNN encoder's hidden states as follows:

$$c_i = \sum_{j=1}^n \alpha_{i,j} h_j \quad (1)$$

where n is the length of the input sequence and the $\alpha_{i,j}$ are weight coefficients. The weight coefficients are computed as follows:

$$\alpha_{i,j} = \frac{\exp(q_{i,j})}{\sum_{k=1}^n \exp(q_{i,k})} \quad (2)$$

$$q_{i,k} = \phi(s_{i-1}, h_k)$$

where the s_i are the hidden states of the decoder and $\phi(\cdot)$ is a feedforward neural network. Unlike the output y_{rnn} of

the RNN encoder-decoder structure, which focuses on the last element of the input sentence, the attention-based contextual vector c_i can take advantage of the semantic information from all words in proportion to their importance.

B. Semantic Features and Image Features

As mentioned earlier, to find the correlation between an image and its corresponding question, representations using feature vectors are needed. The simplest way to generate the semantic features of a question is by averaging the embeddings of the words; however, this approach can yield only coarse estimates of sentence-level features. Currently, it is more common to generate a sentence feature by sequentially passing the embedding of each word through an RNN model and then using the output of the RNN as the sentence embedding. An alternative method is to use an encoder-decoder model consisting of two RNN models, which allows contextual information to be captured by using the hidden states of the encoder and decoder. The decoder can generate an output of flexible length and can be used as either a sequence tagger or a classifier. Unlike semantic features, most image features are extracted using CNNs due to their advantageous ability to extract high-level information from raw two-dimensional (2D) pixel values. Some of the widely used models for this purpose include LeNet, AlexNet, VGG and ResNet [19], [20], [21], [22]. In this paper, we will use a pretrained ResNet model [22] as our image feature extraction model.

C. Multidimensional LSTM

As illustrated in section II-A, the contextual features of a question can be generated using an RNN-based encoder-decoder model. However, since no hidden states exist in a CNN, it is not easy to extract image contextual features in a similar manner. To overcome this obstacle, we use an encoder-decoder structure based on multidimensional LSTM (MDLSTM) [23] to identify contextual features from among the image features generated by a CNN. In our scenario, a 2D MDLSTM network is used to encode and decode image features. Since MDLSTM is not a commonly used LSTM structure, before all of the details are presented, a brief mathematical formulation of MDLSTM is given in (3):

$$\begin{aligned} g^u &= \sigma(W^u H) \\ g_1^f &= \sigma(W_1^f H) \\ g_2^f &= \sigma(W_2^f H) \\ g^o &= \sigma(W^o H) \\ g^c &= \tanh(W^c H) \\ m' &= \sum_{i=1}^2 g_i^f \odot m_i + g^u \odot g^c \\ h' &= \tanh(g^o \odot m') \end{aligned} \quad (3)$$

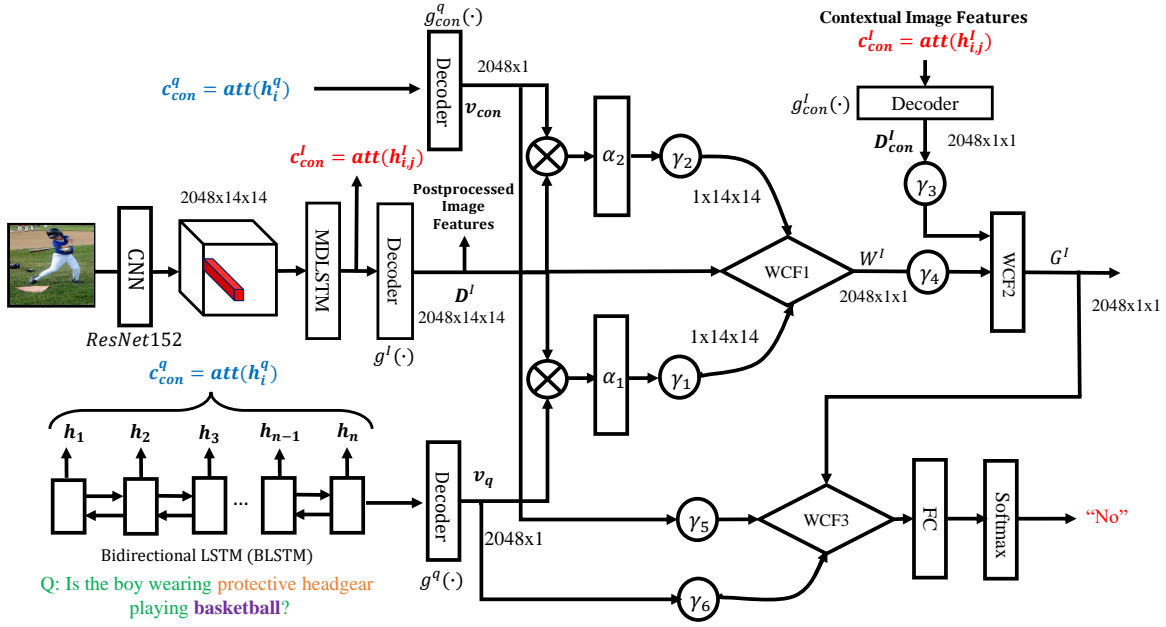


Fig. 2: The general structure of the multimodal attention-based WCF (MA-WCF) model

Here, \mathbf{H} is the concatenation of the new input x_i , transformed by a projection matrix \mathbf{I} , and the hidden vectors \mathbf{h}_i ($i = \{1, 2\}$) from the last time step for both directions in a 2D setup; i.e., $\mathbf{H} = [\mathbf{I}x_i, \mathbf{h}_1, \mathbf{h}_2]^T$. In addition, \mathbf{W}^u , \mathbf{W}_i^f , \mathbf{W}^o , and \mathbf{W}^c are the weight matrices of the input gate, forget gate, output gate and cell state, respectively, in an MDLSTM structure.

From (3), it can be observed that a 2D MDLSTM structure exhibits two main differences from a 1D LSTM structure:

1. Two hidden state vectors \mathbf{h}_i ($i = \{1, 2\}$), one for each of the two directions, from the previous time step are used to generate \mathbf{H} . In our case, the concatenated vector \mathbf{H} for an input image feature $\mathbf{I}_{i,j} \in \mathbf{I}$ is generated from the hidden states corresponding to positions $(i-1, j)$ and $(i, j-1)$, i.e., $\mathbf{h}_{i-1,j}$ and $\mathbf{h}_{i,j-1}$.
2. Two forget gates \mathbf{g}_i^f ($i = \{1, 2\}$), one for each dimension, are used to generate the final output hidden state \mathbf{h}' .

III. ATTENTION-BASED MULTIMODAL VISUAL QUESTION ANSWERING SYSTEM USING WEIGHTED CONTEXTUAL FEATURES

As described earlier, it is possible to extract semantic contextual features using an RNN-based encoder-decoder structure or image contextual features using an MDLSTM-based encoder-decoder structure. Specifically, the RNN structure in our system is chosen to be a bidirectional LSTM (BLSTM) structure [24]. Moreover, as demonstrated by the VQA example given in Figure 1, many instances of misinterpretation in VQA tasks can be attributed to a misunderstanding of the contextual information present in the question (which can be

extremely important) or image. Inspired by these observations and model features, we propose an attention-based multimodal system that leverages contextual features of both questions and images for VQA tasks. The basic structure is shown in Figure 2.

As shown in the figure, our attention-based multimodal system consists of several component:

1. The question contextual feature extraction (Q-CFE) module
 2. The image contextual feature extraction (I-CFE) module
 3. The WCF-based question-image understanding module
 4. The WCF-based answer generation module
- These modules will be described in detail in this section.

A. The Question Contextual Feature Extraction (Q-CFE) Module

A BLSTM-based encoder-decoder structure is used to extract the semantic contextual features in our Q-CFE module. The advantage of the BLSTM structure is that it can capture sentence-level information in both the forward and backward directions; hence, reasonably well-balanced encoded information can be obtained in the last time step n . During training, the question inputs (x_1, \dots, x_n) are read into the BLSTM network in the forward and backward directions, and the network generates two hidden state sequences, $\mathbf{h}\mathbf{f}_t$ and $\mathbf{h}\mathbf{b}_t$. The hidden state \mathbf{h}_t in time step t is then obtained as a concatenation of $\mathbf{h}\mathbf{f}_t$ and $\mathbf{h}\mathbf{b}_t$, i.e., $\mathbf{h}_t = [\mathbf{h}\mathbf{f}_t, \mathbf{h}\mathbf{b}_t]$. For the encoder-decoder structure defined in section II-A, the generated question-level semantic contextual features \mathbf{c}_{con}^q

and the extracted question features v_q can be mathematically represented as follows:

$$\begin{aligned} c_{con}^q &= \sum_{j=1}^n \alpha_{i,j} h_j^q \\ s_t^1 &= g_1(s_{t-1}^1, h_t^q) \\ v^q &= g_1(s_{n-1}^1, h_n^q) \\ s_t^2 &= g_2(s_{t-1}^2, c_{con}^q) \\ v_{con}^q &= g_2(s_{n-1}^2, c_{con}^q) \end{aligned} \quad (4)$$

where $g_1(\cdot)$ is an RNN-based decoder for generating question sequence embeddings and $g_2(\cdot)$ is an RNN-based decoder for generating question-level contextual features, c_{con}^q . s_t^1 is the hidden state generated by $g_1(\cdot)$, and s_t^2 is the hidden state generated by $g_2(\cdot)$. v^q and v_{con}^q are the decoded question features and contextual features, respectively. n is the length of the question, which is equal to the total number of time steps required to process the question. The attention weights $\alpha_{i,j}$ are defined as in section II-A.

B. The Image Contextual Feature Extraction (I-CFE) Module

Following the definition of MDLSTM given in section II-C, a detailed explanation of the construction of the MDLSTM-based I-CFE module is given in this section.

The image feature tensor, containing k image features, is represented by $I \in \mathbb{R}^{k \times n \times n}$. This tensor is extracted in the usual manner, by passing the raw image data into a CNN-based image classifier and then extracting its embedding features before the last softmax layer. Then, the MDLSTM encoder reads in each image feature $I_{i,j} \in I$ together with the hidden state vectors $h_{i,j-1}$ and $h_{i-1,j}$, which are generated from the image features $I_{i,j-1}$ and $I_{i-1,j}$. These hidden states will be used as the input to a subsequent MDLSTM-based decoder to generate decoded image features. By using an attention mechanism similar to that described in section II-A, an image contextual feature is generated as follows:

$$c_{con}^I = att(h_{i,j}^I) = \sum_{i=1}^n \sum_{j=1}^n \alpha_{i,j} h_{i,j}^I \quad (5)$$

Here, $n \times n$ is the number of image features generated by the CNN from the raw image inputs, and $\alpha_{i,j}$ is calculated as follows:

$$\begin{aligned} \alpha_{i,j} &= \frac{\exp(\tau_{i,j})}{\sum_{i=1}^n \sum_{j=1}^n \exp(\tau_{i,j})} \\ \tau_{i,j} &= \phi(s_I, h_{i,j}^I) \end{aligned} \quad (6)$$

where $\phi(\cdot)$ is a feedforward neural network and s_I is the last hidden state generated by the MDLSTM decoder. Since only one feature vector is generated by the decoder for each image, only the last hidden state is applied, i.e., s_I .

The postprocessed image features $D^I \in \mathbb{R}^{k \times n \times n}$ are generated by taking the hidden states from the encoder as the input to the decoder:

$$D_{i,j}^I = g^I(s_{i-1,j}^I, s_{i,j-1}^I, h_{i,j}^I) \quad (7)$$

where $g^I(\cdot)$ is the MDLSTM decoder. Similarly, the decoder output $D_{con}^I \in \mathbb{R}^{k \times 1 \times 1}$ for a contextual feature is calculated as follows:

$$D_{con}^I = g_{con}^I(s_{n-1,n}^I, s_{n,n-1}^I, c_{con}^I) \quad (8)$$

where $s_{n-1,n}^I$ and $s_{n,n-1}^I$ are the previous hidden states of the decoder relative to $s_{n,n}^I$ and c_{con}^I is the contextual image feature generated as shown in (5).

C. The Weighted Contextual Feature (WCF)-based Question-Image Understanding Module

Once the contextual features have been separately extracted from the questions and images, the next important question is how to use these features effectively based on their importance. In this paper, we propose a WCF-based approach to take advantage of the contextual vectors generated by our model. The first application of this technique in solving our problem is to understand the questions asked about the given images, i.e., to find the features of an image that are most closely related to the corresponding image-question pair.

One of the most effective ways to approach this problem is to adopt a (co-)attention mechanism by assigning an attention weight to each image feature vector and then summing them together. There are several popular methods of generating attention weights, including projection [25], HieCoAtt [6], and MCB [10]. In this paper, we will use the projection approach to find the attention weights as shown in (9):

$$\begin{aligned} \alpha_1 &= \sigma((D^I)^T v_q) \\ \sigma(z_j) &= \frac{e^{z_j}}{\sum_{k=1}^n e^{z_k}} \quad \text{for } j = \{1, \dots, n\} \end{aligned} \quad (9)$$

where α_1 is the normalized softmax weighting ($\sigma(\cdot)$) of the inner product of the postprocessed image feature vector D^I and the decoded question feature vector v_q . Since $D^I \in \mathbb{R}^{k \times n \times n}$ and $v_{con} \in \mathbb{R}^{k \times 1 \times 1}$, the attention weights generated by (9) have dimensions of $\alpha_1^T \in \mathbb{R}^{1 \times n \times n}$.

Thus, α_1 represents the attention weights of the image features from the perspective of the question embedding. Similarly, we can generate another attention weight vector α_2 by projecting the question contextual features v_{con}^q onto the postprocessed image feature vector D^I to highlight the image features that are most useful, based on the contextual information contained in the question, by assigning larger attention weights to them:

$$\alpha_2 = \sigma((D^I)^T v_{con}^q) \quad (10)$$

where $\alpha_2^T \in \mathbb{R}^{1 \times n \times n}$.

Although the attention weights generated by using the question embedding v_q and the question contextual features v_{con} have the same dimensions, the naive sum of these two sets of attention weights will not result in suitable performance of our model since the embedding of a sentence and its contextual features may not be of equal importance. This possibility inspires us to assign an additional weight parameter γ_i to each attention weight α_i to enable the further learning of the relative importance of attention weights from different sources.

$$\begin{aligned}\alpha_{1,2} &= \gamma_1 \alpha_1 + \gamma_2 \alpha_2 \\ \mathbf{W}^I &= \alpha_{1,2}^T \mathbf{D}^I\end{aligned}\quad (11)$$

where $\gamma_2 = 1 - \gamma_1$, $0 \leq \gamma_i \leq 1$ ($i = \{1, 2\}$), and \mathbf{W}^I is the attention-generated image feature obtained by using the weighted attention vector $\alpha_{1,2}$. During training, γ_1 is updated through backpropagation.

The image feature \mathbf{W}^I that is generated based on the attention mechanism still does not contain the image contextual features \mathbf{c}_{con}^I generated earlier. To further incorporate this information, we again apply our weighting mechanism to these two vectors as follows:

$$\mathbf{G}^I = \gamma_3 \mathbf{D}_{con}^I + \gamma_4 \mathbf{W}^I \quad (12)$$

where $\gamma_4 = 1 - \gamma_3$ and $0 \leq \gamma_i \leq 1$ ($i = \{3, 4\}$). During training, γ_3 is updated through backpropagation. The generated vector \mathbf{G}^I can be interpreted as a projected image vector based on the extracted understanding of the question and its contextual information.

The next step is to generate an answer based on the question and our projected image vector \mathbf{G}^I .

D. The Weighted Contextual Feature (WCF)-based Answer Generation Module

To generate an answer based on the question-related vectors (v_q and v_{con}) and our generated projected image vector \mathbf{G}^I , a WCF-based convolution algorithm is designed:

$$\begin{aligned}\mathbf{v}_{q,con} &= \gamma_5 \mathbf{v}_{con} + \gamma_6 \mathbf{v}_q \\ \mathbf{A}^{q,I} &= \mathbf{G}^I * \mathbf{v}_{q,con}\end{aligned}\quad (13)$$

where $\mathbf{A}^{q,I}$ is the vector generated by the convolution $\mathbf{G}^I * \mathbf{v}_{q,con}$, $\gamma_5 = 1 - \gamma_6$, and $0 \leq \gamma_i \leq 1$ ($i = \{5, 6\}$). During training, γ_5 is updated through backpropagation.

The convolution operation can be further rewritten as

$$\mathbf{G}^I * \mathbf{v}_{q,con} = FFT^{-1}(FFT(\mathbf{v}_{q,con}) \cdot FFT(\mathbf{G}^I)) \quad (14)$$

Here, $FFT(\cdot)$ denotes the Fourier transform, and FFT^{-1} denotes the inverse Fourier transform.

The final answer is generated in the form of a one-hot vector by passing $\mathbf{A}^{q,I}$ through one fully connected layer and one additional softmax layer.

IV. EXPERIMENT

A. Datasets

MS COCO VQA Datasets: This dataset contains a total of more than 200k images, with 82,783 images for training, 40,504 images for validation, and another 81,434 images for testing. There are 3 questions per image and 10 answers per question. A total of 25% of the test dataset is designated as test-dev data. Currently, there are two versions of the VQA dataset (v1 [26] and v2 [27]), with the same number of images but different numbers of questions. We used both of them in our experiment for completeness, since for most models in the literature, results have been published only for VQA v1.

We report our evaluation results obtained using both the test-standard dataset and the test-dev dataset. Moreover, the results obtained on open-ended tasks (on both VQA v1 and VQA v2) are reported. The model was trained on the training and validation sets, and the results are compared with those of the current state-of-the-art models for each category and the whole dataset.

B. Experimental Setup

1) *Different Model Configurations:* We tested our MA-WCF model with several different configurations:

MA-WCF model without question contextual features: In this model configuration, we removed the question contextual features v_{con} shown in Figure 2; hence, the weight parameters γ_1 , γ_2 , γ_3 and γ_4 were removed from the system during training. The entire model contains only one set of WCFs, i.e., the image contextual features \mathbf{D}_{con}^I .

MA-WCF model without image contextual features: The second model configuration was created by removing the image contextual features from the MA-WCF model, thus eliminating the two weight parameters γ_3 and γ_4 .

MA-WCF model with both question and image contextual features: This is the original model as shown in Figure 2.

2) *Architecture Parameters:* The image features were extracted from a 152-layer ResNet model [22] that was pre-trained on the ImageNet dataset [32]. The features extracted before the last classifier layer ("pool5") were used as the image feature inputs to the MDLSTM classifier. Following the model configuration described in [10], image features were generated with dimensions of $I \in \mathbb{R}^{2048 \times 14 \times 14}$. The dimensionality of the MDLSTM encoder-decoder structure was set to 2, and the number of units was chosen to be 2048 to keep the dimensions consistent. On the question side, each question was first tokenized into words, and 100-dimensional word vectors were generated from GloVe word2vec representations [33]. Then, these word vectors were fed into a 2048-unit BLSTM structure. The decoder RNN had the same number of units, such that $v_q \in \mathbb{R}^{2048 \times 1}$. The other vectors' dimensional information can be found directly from Figure 2.

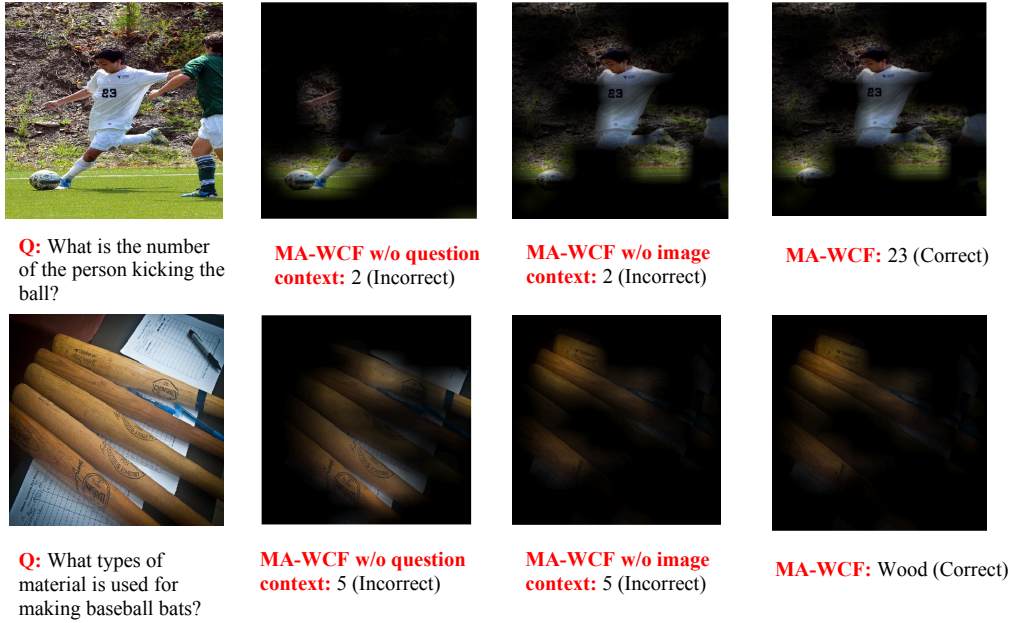


Fig. 3: Examples used for comparison between MA-WCF and state-of-the-art models

Table I: Results for Open-ended Answers on the Test-dev and Test-standard Datasets in VQA v1 and VQA v2

VQA v1	Test-dev (%)				Test-standard (%)			
	Y/N	Num.	Other	All Categories	Y/N	Num.	Other	All Categories
Ensemble MCB [10]	83.4	39.8	58.5	66.7	83.2	39.5	58.0	66.5
Stacking+Features [5]	NA	NA	NA	NA	82.6	39.5	58.3	67.3
Alpha VQA [28]	NA	NA	NA	NA	87.61	45.63	63.30	71.48
MA-WCF w/o Question Context	84.63	43.51	61.32	69.27	84.73	43.13	63.56	69.3
MA-WCF w/o Image Context	85.13	43.88	62.83	69.56	85.3	44.13	63.39	69.7
MA-WCF	88.64	46.52	64.12	72.45	88.92	46.73	64.46	73.52
VQA v2	Test-dev (%)				Test-standard (%)			
	Y/N	Num.	Other	All Categories	Y/N	Num.	Other	All Categories
IL-QTA [29]	87.96	56.12	63.51	72.75	88.26	55.22	63.63	72.93
MIL@HDU [30]	90.09	59.2	65.69	75	90.36	59.17	65.75	75.23
GridFeat [31]	90.73	61.84	67.01	76.19	90.81	61.53	67.04	76.29
MA-WCF w/o Question Context	89.62	58.28	65.72	75.95	89.45	58.46	66.58	75.92
MA-WCF w/o Image Context	89.95	59.45	66.43	76.25	90.35	59.82	67.34	76.27
MA-WCF	91.24	61.05	67.43	76.94	91.45	61.32	67.85	77.05

The Adam stochastic optimizer was used, with a learning rate of 0.002 and early termination in the case of no improvement for 30 consecutive epochs.

C. Experiments

The three different configurations of our model were trained on the training and validation sets from the MS COCO VQA dataset for comparison with state-of-the-art models. The results are shown in Table I. In this subsection, we present some visual examples showing how the MA-WCF structure

facilitates interpretation by taking advantage of both question and image contextual features.

1) Interpretability of MA-WCF: Figure 3 shows the results of applying the three different MA-WCF configurations to two different examples. In the first example, the noun “the person” is modified by the postpositional phrase “kicking the ball”; therefore, question-level contextual features must be considered for the question to be correctly understood. Otherwise, the model may not be able to focus on the correct subregions of the image since the question is not fully

understood. In the second configuration, with the image-level contextual features removed, the model can locate the correct subregions of the image; however, it still cannot generate the correct answer since the contextual correlations between the masked sub-regions cannot be fully understood. Therefore, incorrect answers are generated when the image contextual features are not considered.

Similarly, in the second example, the question-level semantic contextual features help to locate the correct image subregion(s) to enable the identification of the “type of material”. Then, the image contextual features further help to generate the correct answer by filtering out some noisy image information (such as the presence of 5 baseball bats) by giving them lower weights, hence generating the correct answer.

From the above two examples, we can see that the question-level semantic contextual features help the system to locate the correct question-related subregions of the image (through masking), while the image-level contextual features mainly help the system to generate correct answers by assigning appropriate weights to the extracted subregions.

In the next subsection, we will present more quantitative results obtained using our model and compare them with the results of several baseline models on the VQA datasets.

2) *Experiment Results on VQA Datasets*: One observation that can be drawn from Table I is that the performance of the model without question contextual features is far inferior to both that of the model without image contextual features and that of the model with both types of contextual features. This finding demonstrates the importance of question contextual features to our system. The MA-WCF model with both image and question contextual features outperforms the previous state-of-the-art results on each category for both the test-dev and test-standard sets in VQA v1 and on most categories in VQA v2. Excitingly, our model even shows better performance than ensemble/stacking-based models do. On VQA v1, the MA-WCF model outperforms the current state-of-the-art Alpha VQA model by 2.1% on the test-standard dataset. On VQA v2, the MA-WCF model outperforms the current state-of-the-art model by GridFeat by 0.75% on the test-dev dataset and by 0.76% on the test-standard dataset.

V. CONCLUSION

In this paper, we have proposed a novel interpretable multimodal system using attention-based weighted contextual features (MA-WCF) to address the visual question answering (VQA) problem. By using adaptively weighted contextual features extracted from both questions and images, our system gains the advantageous ability to pinpoint the most important parts of both questions and images while de-emphasizing less important features. We have achieved new state-of-the-art results on the MS COCO VQA dataset for open-ended

question tasks. As a relatively general technique, our MA-WCF approach can be further extended to other text- or image-related tasks, such as question answering, text summarization or visual grounding. The interpretability of the model also endows it with great potential for application to more complex conversational VQA tasks; we are currently working on this problem and will report our progress in future works.

REFERENCES

- [1] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, “Devise: A deep visual-semantic embedding model,” in *Advances in neural information processing systems*, 2013, pp. 2121–2129.
- [2] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, “Improving image-sentence embeddings using large weakly annotated photo collections,” in *European Conference on Computer Vision*. Springer, 2014, pp. 529–545.
- [3] M. Malinowski, M. Rohrbach, and M. Fritz, “Ask your neurons: A neural-based approach to answering questions about images,” in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2015, pp. 1–9.
- [4] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.
- [5] N. F. Rajani and R. J. Mooney, “Stacking with auxiliary features,” *arXiv preprint arXiv:1605.08764*, 2016.
- [6] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.
- [7] D. Teney, P. Anderson, X. He, and A. v. d. Hengel, “Tips and tricks for visual question answering: Learnings from the 2017 challenge,” *arXiv preprint arXiv:1708.02711*, 2017.
- [8] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, “Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,” *IEEE Transactions on Neural Networks and Learning Systems*, 2018.
- [9] Y. Zhang, J. Hare, and A. Prügel-Bennett, “Learning to count objects in natural images for visual question answering,” *arXiv preprint arXiv:1802.05766*, 2018.
- [10] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 457–468.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [12] Z. S. Harris, “Distributional structure,” *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.
- [13] R. Cadene, H. Ben-Younes, N. Thome, and M. Cord, “Murel: Multimodal Relational Reasoning for Visual Question Answering,” in *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2019. [Online]. Available: <https://arxiv.org/abs/1902.09487>
- [14] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015.
- [15] B. Liu and I. Lane, “Attention-based recurrent neural network models for joint intent detection and slot filling,” *arXiv preprint arXiv:1609.01454*, 2016.
- [16] Y. Wang, M. Huang, L. Zhao *et al.*, “Attention-based lstm for aspect-level sentiment classification,” in *Proceedings of the 2016 Conference*
- [17] L. Itti and C. Koch, “Computational modelling of visual attention,” *Nature reviews neuroscience*, vol. 2, no. 3, p. 194, 2001.
- [18] 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 *International Conference on Machine Learning*, 2015, pp. 2048–2057.
- [19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
- [20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [22] A. Graves and J. Schmidhuber, “Offline handwriting recognition with multidimensional recurrent neural networks,” in *Advances in neural information processing systems*, 2009, pp. 545–552.
- [23] —, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
- [24] 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*, 2016, pp. 4613–4621.
- [25] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual Question Answering,” in *International Conference on Computer Vision (ICCV)*, 2015.
- [26] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [27] “Alphavqa,” URL: <https://visualqa.org/roe.html>, 2016.
- [28] D. Tuong, T. Huy, D. Thanh-Toan, T. Erman, and D. T. Quang, “Interaction learning with question-type awareness for visual question answering,” in *Computer Vision and Pattern Recognition, 2019. CVPR 2019. IEEE Conference on*. IEEE, 2019.
- [29] “State-of-the-art vqa model in 2019 vqa challenge,” URL: <https://visualqa.org/roe.html>, 2019.
- [30] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, “In defense of grid features for visual question answering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 267–10 276.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
- [32] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.