Rich Visual Knowledge-Based Augmentation Network for Visual Question Answering

Liyang Zhang, Shuaicheng Liu, *Member, IEEE*, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, *Senior Member, IEEE*, and Lianli Gao, *Member, IEEE*

***Abstract*— Visual question answering (VQA) that involves understanding an image and paired questions develops very quickly with the boost of deep learning in relevant research fields, such as natural language processing and computer vision. Existing works highly rely on the knowledge of the data set. However, some questions require more professional cues other than the data set knowledge to answer questions correctly. To address such an issue, we propose a novel framework named a knowledge-based augmentation network (KAN) for VQA. We introduce object-related open-domain knowledge to assist the question answering. Concretely, we extract more visual information from images and introduce a knowledge graph to provide the necessary common sense or experience for the reasoning process. For these two augmented inputs, we design an attention module that can adjust itself according to the specific questions, such that the importance of external knowledge against detected objects can be balanced adaptively. Extensive experiments show that our KAN achieves state-of-the-art performance on three challenging VQA data sets, i.e., VQA v2, VQA-CP v2, and FVQA. In addition, our open-domain knowledge is also beneficial to VQA baselines. Code is available at https://github.com/yyyanglz/KAN.**

***Index Terms*— Knowledge base, object detection, self-attention, visual question answering (VQA).**

I. Introduction

VISUAL question answering (VQA) [1]–[4] is a cross-modality task that combines computer vision (CV) [5], [6] and natural language processing (NLP) [7], [8]. Such a cross-modality task requires understanding not only the content of visual inputs but also the contextual knowledge expressed by questions. In particular, given a picture and a question based on the picture in natural language, the VQA task needs to integrate both visual features and semantic context to give a correct answer. However, such a task requires a comprehensive understanding of both visual and linguistic components. The right answer can be produced only when both of them infer correctly.

For multimodality feature representation, previous works get accustomed to use visual feature extracted by convolutional neural network (CNN) [42], e.g., VGG [9] and ResNet [6]. With further development, Anderson *et al.* [3] proposed a novel feature extraction method in vision and language tasks, which can extract more representative features to promote the performance. Especially, they used an object detection model, Faster R-CNN, to detect instances of objects belonging to certain classes and localize them with corresponding coordinates in bounding boxes. Also, they used the method to win first place in the 2017 VQA Challenge. On the other hand, the semantic feature of questions is achieved with word embedding, which was pretrained [10]. The role of word embedding is to convert every single word to a high dimension vector, where semantic similarities are represented by distances. To get the contextual feature of the whole question, the set of the word embeddings is then sent into a recurrent neural network (RNN) [11], [46].

Even though VQA has drawn plenty of attention these years, some critical challenges still need to be solved. For multimodality feature fusion, most existing methods focus on the fusion with different modalities, i.e., visual features from images and semantic features from questions. The typically used fusion method, attention mechanism, is utilized to obtain the corresponding relationships between object regions on images and words from questions [43]. For instance, previous work coattention [12] learns the most relevant relations between object region and question word pairs to obtain the right answers. Also, some works explore the relation in a single modality, e.g., language modality. The BERT model [8] grabs the relations between word-to-word pairs from the question by the self-attention mechanism. When the fusion mechanism deals with two or more kinds of modalities, the current methods struggle to balance the impact among these kinds of modality features.

In addition, for image inputs, there are no links or relationships between the object features extracted by Faster R-CNN.

Manuscript received December 7, 2019; revised May 31, 2020; accepted August 6, 2020. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2019J073, in part by the National Natural Science Foundation of China under Grant 61772116, Grant 61872064, Grant 61632007, Grant 61602049, and Grant 61872067; in part by the Sichuan Science and Technology Program under Grant 2019JDTD0005 and Grant 2019YFH0016, in part by the Open Project of Zhejiang Lab under Grant 2019KD0AB05, and in part by the Zhejiang Lab's International Talent Fund for Young Professionals. *(Corresponding author: Lianli Gao.)*

Liyang Zhang, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao are with the Future Media Center, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: lianli.gao@uestc.edu.cn).

Shuaicheng Liu is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with Megvii Technology Ltd., Chengdu 611730, China.

Donghao Liu is with Megvii Technology Ltd., Chengdu 611730, China.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3017530

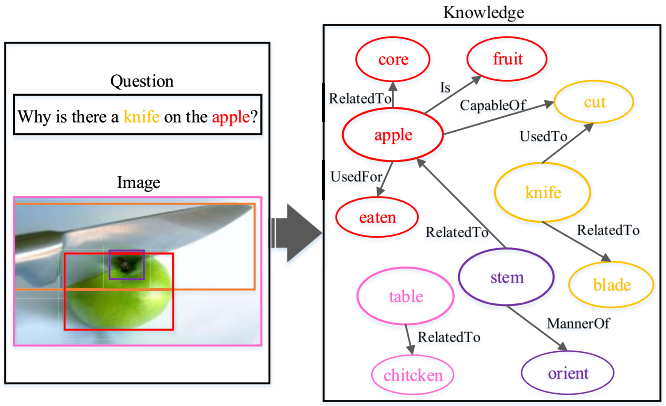


Fig. 1. Object words of questions and images are used to search related knowledge from ConceptNet [16]. For example, the queried knowledge of the object words "apple" and "knife" is related closely. The related knowledge is also helpful to answer the question "Why is there a knife on the apple?"

Previous methods, such as graph network [13]–[15], can capture the relationship between objects. However, the graph network is hard to converge in the training stage, and the current performance of these methods is not as good as expected. Moreover, for some questions concerned with logical reasoning, errors in answers occur as images only hold visual information and contain no relevant knowledge.

To solve such problems, we propose a novel knowledge-based augmentation network (KAN) structure that can balance the weight of extra knowledge and detected objects adaptively. First, the additional visual feature is extracted from the image to inform our model with more tiny objects, some of which barely draw any attention in previous methods. Second, we simulate the deposit of knowledge in human brains so that we provide necessary external knowledge from a large-scale knowledge base ConceptNet [16]. ConceptNet, which contains links between entities and the related facts, compensates for extraneous data for images that may alone lack sufficient information. Different from the knowledge base in [17], we employ ConceptNet to obtain external knowledge with labeled relationships and reliability scores. It is not proper to treat visual features and knowledge information indiscriminately when we integrate them to obtain the answer. Consequently, we carefully design an adaptive score attention module that, for different questions, it automatically adjusts the importance of information feeding to the model between rich visual cues and the extra knowledge base. Especially, for a question on logical reasoning, the adaptive score attention module allocates more weights to external knowledge. However, for a question concerned about the color or number of specific objects on the image, the module prefers to assign more attention to visual features. Through the adaptive score attention module, our model leverages images with external knowledge by a more exhausted mode so that it delivers better accuracy upon some special questions. Our contributions are summarized as follows:

1) We introduce an external knowledge base to enrich the training knowledge base, which provides a more common sense that is absent in the training knowledge base. It helps to answer some specific and professional questions about reasoning based on facts or experience.
2) We adopt adaptive score attention, which can automatically choose whether we use external knowledge or current image representations, to assist in answering the question.
3) Extensive experiments are conducted on three challenging benchmarks (VQA v2, VQA-CP v2, and FVQA), and experiment results show that our approach achieves state-of-the-art performance. The effect of importance is also well exploited in the ablation study.

II. Related Works

A. Visual Question Answering

In the early stage of VQA research [2], [18], [19], a CNN is usually utilized to extract global features of an image, and the corresponding question is fed into LSTM networks [20] to prepare contextual features. Cross-modality models [21], [22], on the other hand, try to combine these two inputs for the question answering. Visual features are composed of multiple grids that have the same shapes and sizes with each grid cell containing only partial data of objects [23]. Therefore, spatial grid features can cause loss of visual information inevitably. Anderson *et al.* [3] proposed a bottom–up attention mechanism that extracts features on the level of object region from an image, which successfully maintains the whole objects information. However, the visual features on object regions are still not fully utilized as they lack internal links.

B. Self-Attention

The attention mechanism [7], [24] is early employed in NLP tasks and then adopted in CV fields [25]. Vaswani *et al.* [26] proposed a self-attention approach based on the original attention so that they can weight different positions in a sequence with different importances. Yu *et al.* [27] leveraged the self-attention mechanism to deal with correlations among different object regions and relationships between different word sequences. Yu *et al.* [28] combined self-attention to propose a new method in the field of dense video captioning that can use nonrecurrent structure to encode generated sentences to improve the performance.

C. External Knowledge Base

The external knowledge has been wildly adopted not only in the NLP [29]–[31] but also in the field of CV tasks [32]. Many large-scale common knowledge bases are available, such as YAGO2 [33], DBpedia [34], and ConceptNet [16]. Wu *et al.* [17] extracted knowledge relevant to the corresponding image from DBpedia and directly feed external knowledge combined with visual features into their model. External knowledge that has connections with objects in an image is also extracted by Gu *et al.* [35] from ConceptNet and is combined with the image to generate a scene graph.

Different from the works in [17] and [32], we employ ConceptNet as our external knowledge base in our work, which is a freely available knowledge base. ConceptNet provides various

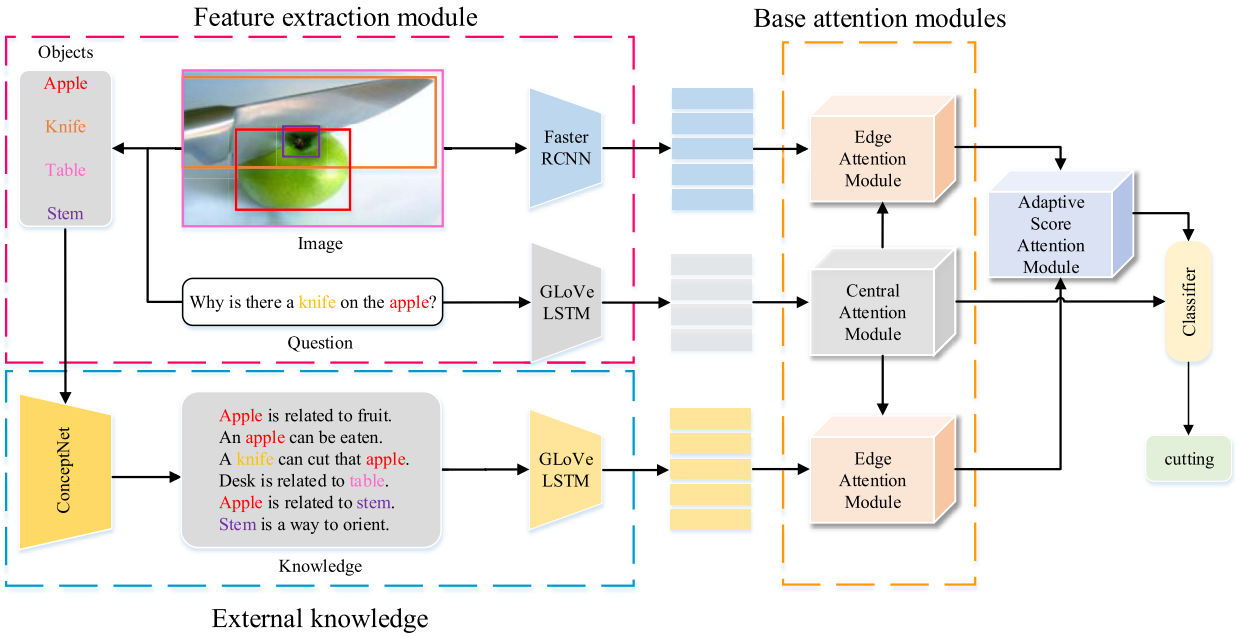


Fig. 2. Framework of the whole KAN. First, the question feature processed by GloVe word embedding and LSTM passes through the CAM to get the attended question feature. Then, the attended question feature and visual feature from images pass through the top EAM to obtain the attended visual feature. External knowledge is queried through ConceptNet, and the attended knowledge feature is obtained by the bottom EAM as earlier. The attended visual feature and the attended knowledge feature get balanced attention by the impact to the question through the adaptive score attention module. The balanced knowledge–image feature and semantic feature of questions are sent into the classifier to gain the final answer.

labeled relationships to connect entities, and the knowledge is connected with these labeled relationships to present capability and features of entities. In ConceptNet, a score is attached to measure the reliability degree of knowledge. In addition, we design a filtering method to pick valuable knowledge with the reliability score. Concretely, we sort the queried knowledge of entities and only obtain the knowledge with a high reality score. However, DBpedia, which is the knowledge base employed in [17], has no features mentioned earlier. DBpedia fetches structural data based on Wikipedia terms with richer information but less focus.

In addition, we design a separate module called the knowledge edge attention module (KEAM). VQA is a task that answers a specific question, and during the process of inferencing, we utilize the single module KEAM to deal with the relationship between external knowledge and question, which can extract principal information from external knowledge effectively and specifically. In [17] and [32], there is no single module to handle external knowledge and only one module to deal with the relationship between external knowledge, image, and question.

III. Our Method

In this article, we aim to efficiently extract rich knowledge features and then improve the fusion of knowledge and object representation to provide an accurate answer. Our proposed framework (KAN) is shown in Fig. 2, which consists of: 1) a feature extraction module that extracts detailed image feature and question feature; 2) external knowledge that provides a more common sense that is absent in the training image and question pairs; 3) base attention modules that include central attention and edge attention module (EAM); and 4) adaptive score attention module that balances the weight of extra knowledge and detected objects. In the following, we present the details of the abovementioned four major components.

A. Feature Extraction Module

In this section, we introduce the first component, namely, the feature extraction module, which obtains two kinds of features.

1) Image Feature: The previous method extracted the adaptive number of object regions from each image, which may cause the loss of some key information [3]. The information of the object region below an adaptive threshold holds beneficial knowledge for the model. Here, we adopt the fixed-size object region features. Especially, we use Faster R-CNN [36] pretrained on the Visual Genome data set to obtain the representations of an input image I by extracting fixed-M object-based region vectors. This process can be formulated as follows:

$$\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_M}] = f(I), \quad \mathbf{X}_i \in R^m \tag{1}$$

where X_i represents the ith object proposal feature, $f(*)$ indicates the Faster R-CNN model, M is the number of the object-based regions, and m stands for the dimension of each object feature. Image features are represented as $\mathbf{X} \in R^{m \times M}$, $m = 2048$, and $M = 100$.

2) Question Feature: Given a question $Q = [q_1, q_2, \ldots, q_N]$, we first transform the question Q into a lower dimension feature $\hat{\mathbf{Q}}$ with GloVe word embedding [10], [37]. Then, we employ a single layer long short-term memory (LSTM)

to encode the word embedding $\hat{\mathbf{Q}}$. For the ith step, the output of hidden state is denoted as $\mathbf{h}_i$. To get the comprehensive word-level question information, we obtain all the hidden state with dimension n generated by the LSTM. Thus, the question feature can be represented as $\mathbf{Y} \in R^{n \times N}$, $n = 512$, and $N = 14$. We formulate this process as follows:

$$\hat{\mathbf{Q}} = \text{Glove}(\mathbf{Q}) = [\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots, \hat{\mathbf{q}}_N] \tag{2}$$

$$\mathbf{h}_i = \text{LSTM}(\hat{\mathbf{q}}_i, \mathbf{h}_{i-1}), \quad i \in [1, N] \tag{3}$$

$$\mathbf{Y} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]. \tag{4}$$

B. External Knowledge

1) Attribute and Object Extractor: While object region features are gathered from images, Faster R-CNN is also deployed to provide attribute labels and object labels for the object region features

$$\text{Attr} = [\text{Attr}_1, \text{Attr}_2, \ldots, \text{Attr}_M] = f_a(I) \tag{5}$$

$$\text{Obj} = [\text{Obj}_1, \text{Obj}_2, \ldots, \text{Obj}_M] = f_o(I) \tag{6}$$

where Attr and Obj represent attribute labels and object labels of images, respectively.

Object labels (e.g., window, banana, and sky) and attribute labels (e.g., green, red, and wooden) have separate influences on a model. Attributes provide decorations to objects, but the presence of attributes, sometimes, confuses the inference of a model. For example, if a question examines "what color the t-shirt is?" by confronting "green" and "red" among processed attribute labels, then the model is surrounded by unnecessary noise with a high probability of confounding. Thus, we drop the attributes that may distract a model to questions. Moreover, most questions in VQA are relevant to objects in images. If a question asks "Do both elephants have tusks?" then the question concerns the real elephants in a certain image, which requires some knowledge of elephants.

As most object labels from images are unrelated to questions, noise object labels must be rejected by selection according to the following two steps.

1) Only object labels that are present in the corresponding questions are extracted. These object labels, annotated as O_{ex}, explicitly correlate to questions such that they are serving to directly provide correct answers.
2) Although the most frequent object labels have no direct relationships with the questions, they are also collected to provide implicit connections to other objects inside the questions, annotated as O_{im}.

As such, final object labels are composed of both explicit and implicit object labels

$$O = \{O_{\text{ex}}, O_{\text{im}}\}. \tag{7}$$

2) Getting External Knowledge of Object Labels: To answer questions involving images that lack the necessary common sense, we leverage a large-scale knowledge graph from ConceptNet with massive contextual information between real objects inside. To be more precise, we query the knowledge from both explicit object labels O_{ex} and implicit object labels O_{im} in ConceptNet. The knowledge includes the following segments:

$$\text{knowledge} = \{\text{fact}, w\} \tag{8}$$

where fact and w are the external knowledge and the corresponding weight of an object label, respectively. One piece of knowledge fact represents a descriptive statement, which describes the attribute, common sense, or capability about an entity word. Weight w stands for the degree that we can rely on the corresponding piece of knowledge; a higher w value means more confident dependence on the specific knowledge

$$\text{fact} = \{O, R\} \tag{9}$$

where fact and R represent the knowledge and the corresponding relationship against the object entity O, respectively. "fact" is a sentence in a natural language format composed of a relationship linking the object entity. To be concrete, the fact also can be represented in natural language format as

$$\text{fact} = [s_1, s_2, \ldots, s_L] \tag{10}$$

where s_i is a single word in natural language.

Sentences in natural language format are discrete sequences and, thus, cannot be directly utilized by the inference in our model. To solve such a problem, word embedding is applied to every single word within a sentence of knowledge to map them into a continuous high-dimensional space. The linguistic information is extracted as

$$\mathbf{S}_l = \mathbf{W}_s s_l, \quad l \in [1, L] \tag{11}$$

where $\mathbf{W}_s$ are learnable parameters and L is the length of the sentence. After processed by word embedding, knowledge only contains linguistic information but lacks contextual relationship. LSTM is deployed to generate context data relevance

$$\mathbf{h}_l = \text{LSTM}(\mathbf{S}_l, \mathbf{h}_{l-1}) \tag{12}$$

where $\mathbf{h}_l$ is the hidden state of LSTM.

Finally, we sufficiently use each hidden state of $\mathbf{h}_l$ instead of taking the last hidden state $\mathbf{h}_L$ to represent the knowledge fact of the object

$$\mathbf{F} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L] \tag{13}$$

where $\mathbf{F}$ is a high-dimensional matrix, indicating the semantic and contextual features of the knowledge.

C. Attention Module

1) Base Attention Unit: For various inputs, a detailed image, a question, and the external knowledge, we design the attention module to create a model that handles multiple sources of data based on a cross-modality model called modular coattention networks (MCANs) [27]. Our model is a symmetric modular structure consisting of two sections: 1) one central attention module (CAM) and 2) two edge-attention modules that are shown in Fig. 3.

Both the CAM and the edge-attention module are stacked with scaled dot-product attention [26]; thus, we first introduce this fundamental module. The vector inputs of each layer for scaled dot-product attention are duplicates of queries,

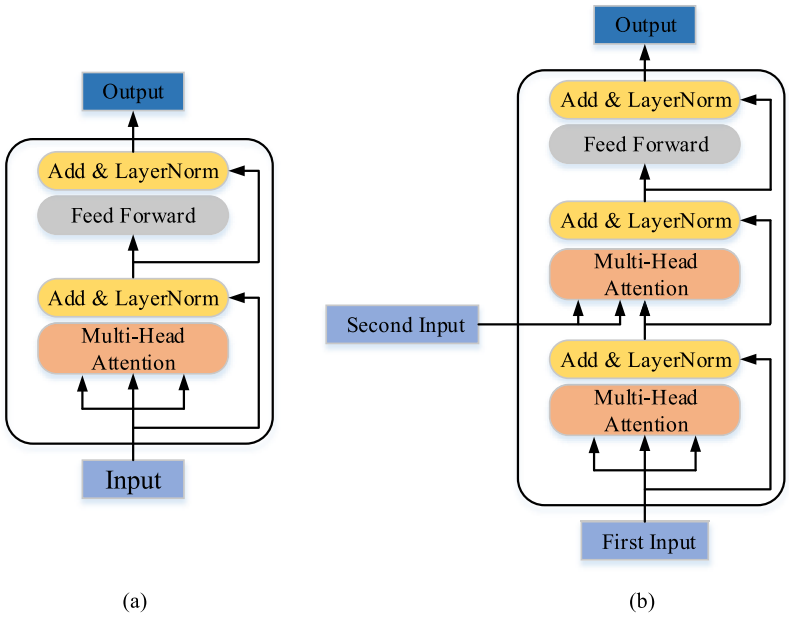


Fig. 3. Two main attention modules in our base attention modules. Each attention module is composed of the scaled dot-product attention. (a) CAM takes the question feature as input and outputs the attended question feature. (b) EAM takes the visual feature or knowledge feature as the first input and question feature as the second input while outputs the attended visual feature or attended knowledge feature, respectively.

keys, and values, and their dimensions are d_q, d_k, and d_v, respectively. For computational convenience, we let $d_q = d_k = d_v = d$ and pack different queries, keys, and values together into matrix $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$.

As a result, we transform the dot-product operation between queries and keys into the multiplication of matrix $\mathbf{Q}$ and matrix $\mathbf{K}$ followed by dividing a scaling factor, i.e., the value of the square root of d_k. Then, previous output is passed to the softmax function, following which the attention weights are obtained by multiplying matrix $\mathbf{V}$:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_k}}\right)\mathbf{V}. \tag{14}$$

Furthermore, to improve the expressive capability of the attended features, we introduce multihead attention to collect information from various feature subspaces

$$\text{MH}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(\mathbf{head_1}, \ldots, \mathbf{head_h})\mathbf{W^O} \tag{15}$$

$$\mathbf{head_t} = \text{Attention}\left(\mathbf{Q}\mathbf{W_t^Q}, \mathbf{K}\mathbf{W_t^K}, \mathbf{V}\mathbf{W_t^V}\right) \tag{16}$$

where $\mathbf{W^O} \in R^{d\times d}$. $\mathbf{W_t^Q}, \mathbf{W_t^K}, \mathbf{W_t^V} \in R^{d\times d_h}$, $d_h = d/h$, d is the dimension of the query, key, and value, and h is the number of the heads mentioned earlier. $\mathbf{W_t^Q}$, $\mathbf{W_t^K}$, and $\mathbf{W_t^V}$ are the learnable mapped matrices of the queries, the keys, and the values of the tth $\mathbf{head_t}$, respectively.

2) Central Attention Module: Each layer of CAM is stacked by G layers scaled dot-product attention. Its input question feature is formulated as follows:

$$\mathbf{Y} = [\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_N}] \in R^{n\times N}. \tag{17}$$

Then, $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{Y}$, as shown in Fig. 3(left). A softmax function is added to each layer's output of the scaled dot-product attention to learn the attention weight between arbitrary question word y_i and y_j, and an attention matrix is obtained. Next, we apply the attention matrix to the question feature to produce the self-attended question feature $\mathbf{Y^g}$, as shown in (14), which is used as the input of the next scaled dot-product attention layer. This process is formulated as

$$\mathbf{Y^g} = \text{CAM}^g(\mathbf{Y^{g-1}}) \tag{18}$$

where the initial input $\mathbf{Y^0} = \mathbf{Y}$ and each CAM is a scaled dot-product attention. The final output of the CAM is $\mathbf{Y^G}$.

3) Edge Attention Module: EAM is merged by Image EAM (IEAM) and KEAM. IEAM and KEAM share the same model structure where each contains G stacked layers that are composed of two layers of scaled dot-product attention. Take IEAM for example, the input of IEAM coming from the detailed visual feature $\mathbf{X}$ mentioned earlier plus the final output named attended question feature $\mathbf{Y^G}$ from the CAM. The first scaled dot-product attention in each layer of IEAM with the same function as the single-layer CAM computes self-attention of visual feature input $\mathbf{X}$.

The second scaled dot-product attention is operated by a softmax function to learn an attention weight matrix that describes the correlation between the object region feature x_i and the question word feature y_j. Consequently, the attention matrix is applied to the attended question feature $\mathbf{Y^G}$ from the last layer of CAM to output the question-based attended image feature $\mathbf{X^g}$, which is the input of the next layer of the scaled dot-product attention. The pipeline is

$$\mathbf{X^g} = \text{IEAM}^g(\mathbf{X^{g-1}}, \mathbf{Y^G}) \tag{19}$$

where the initial input $\mathbf{X^0} = \mathbf{X}$, and $\mathbf{Y^G}$ is the output of the final layer from the CAM. The final G layers' output of IEAM is $\mathbf{X^G}$, which represents the attended image feature.

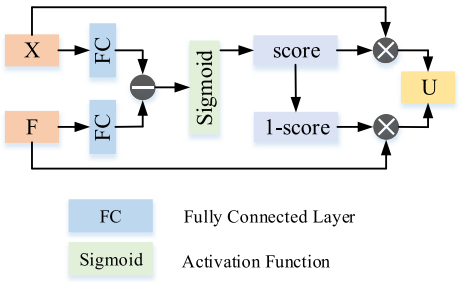


Fig. 4. Adaptive score attention module. It balances the impact of attended visual feature $\mathbf{X}$ and attended knowledge feature $\mathbf{F}$ according to the specific question.

Similarly, the correlation between the external knowledge $\mathbf{F}$ and the attended question feature $\mathbf{Y^G}$ is calculated through G layers of KEAM

$$\mathbf{F^g} = \text{KEAM}^g(\mathbf{F^{g-1}}, \mathbf{Y^G}) \tag{20}$$

where the initial input $\mathbf{F^0} = \mathbf{F}$, and $\mathbf{Y^G}$ is the output of the final layer from the CAM. The final output $\mathbf{F^G}$ of the G layers in KEAM is the attended knowledge feature.

D. Adaptive Score Attention Module

The detailed visual feature $\mathbf{X}$, the external knowledge $\mathbf{F}$, and the question feature Y are processed through our symmetric modular model to compute the attended visual feature $\mathbf{X}^G$, the attended knowledge feature $\mathbf{F}^G$, and the attended question feature $\mathbf{Y}^G$.

When coming across different questions, the attended visual feature $\mathbf{X}^G$ and the attended knowledge feature $\mathbf{F}^G$ have flexible impacts at providing the correct answer to the specific question. Indiscriminately utilizing these two source information leads to a universal standard to extract hidden information, thus a less exploit of rich visual information with the external knowledge base. If a question focuses on the object and scene in the image, for example, a question like "What color is the umbrella?", the detailed visual feature is expected to present a correct answer. However, if a question needs more knowledge to infer the right answer, say a question asks "Why is the man on the street?" a more reasonable answer such as "homeless" can be indicated with the guidance of extraneous knowledge. Therefore, for different questions, visual feature and knowledge feature play distinct roles.

Therefore, we carefully design an adaptive score attention module, that to a specific question, automatically picks one source information that is more suitable for providing an accurate answer while treating the other as an auxiliary, as shown in Fig. 4. First, we design a score function to compute the scores for the detailed visual feature $\mathbf{X}$ and the external knowledge $\mathbf{F}$

$$S(\mathbf{X}) = \mathbf{W_2^X}\left(\tanh\left(\mathbf{W_1^X X^G}\right)\right) \tag{21}$$

$$S(\mathbf{F}) = \mathbf{W_2^F}\left(\tanh\left(\mathbf{W_1^F F^G}\right)\right) \tag{22}$$

where $\mathbf{W_1^X} \in R^{o\times m}$, $\mathbf{W_2^X} \in R^{o\times o}$, $\mathbf{W_1^F} \in R^{o\times n}$, $\mathbf{W_2^F} \in R^{o\times o}$, and all are learnable parameters. $S(\mathbf{X})$ and $S(\mathbf{F})$ are indicators to depict how important the detailed image feature. The external knowledge is

$$\mathbf{A_X} = \sigma\left(S(\mathbf{X}) - S(\mathbf{F})\right) \tag{23}$$

$$\mathbf{A_F} = 1 - \mathbf{A_X} \tag{24}$$

$$\mathbf{U} = \mathbf{A_X X} + \mathbf{A_F F} \tag{25}$$

where σ is the sigmoid activation function, and $\mathbf{A_X}$ is the adaptive attention of detailed image feature $\mathbf{X^G}$, which is computed by comparing the impacts of visual feature over knowledge feature. $\mathbf{A_F}$ is the adaptive attention of external knowledge $\mathbf{F^G}$. $\mathbf{U}$ represents the merged fusion feature from visual feature and knowledge feature by the adaptive attention fed by $\mathbf{A_X}$ combined with $\mathbf{A_F}$.

TABLE I

VQA v2, VQA-CP v2, AND FVQA DATA SET STATISTICS. "#IMAGE" DENOTES THE NUMBER OF IMAGES IN THE SPLITS, AND "#QA PAIR" INDICATES THE NUMBER OF QUESTION–ANSWER PAIRS IN THE CORRESPONDING SPLIT

Dataset	Split	#Image	#QA pair
VQA v2	Train	80k	444k
	Validation	40k	214k
	Test	80k	448k
VQA-CP v2	Train	121K	438K
	Test	98K	220K
FVQA	Train	1100	2927
	Test	1090	2899

IV. EXPERIMENTS

In this section, we train our model on the large VQA data set VQA v2 [38], newly proposed data set VQA-CP v2 [39], and knowledge-based data set FVQA [40]. Moreover, various experiments are conducted on the data set to validate the effectiveness of our model. The results are compared with previous state-of-the-art methods quantitatively and qualitatively.

A. Data Sets

1) VQA v2: VQA v2 is the most popularly adopted data set of VQA tasks. It consists of three types of questions: Yes/No, Number, and Other. 1106k human-annotated question–answer pairs are included in the VQA v2 data set, within which 204k images come from Microsoft COCO data set. Every image has three questions at least and 5.4 questions on average. Each question has ten ground-truth answers annotated by ten different people, where people who provide answers are not the same as people who ask questions. The answers are evaluated with the accuracy metric as follows [1]:

$$\text{accuracy} = \min\left(\frac{\#\text{humans provided answer}}{3}, 1\right) \tag{26}$$

where #humans provided answer means the number of humans that provided the answer.

The data set is split into the training set, validation set, and testing set, as shown in Table I. Besides, 25% of the test set is sampled as the test-dev data set to evaluate the performance of the model online, while a 100% test set is denoted as the test-std data set to assess the performance of the model comprehensively. All the results are reported based on both the test-std and test-dev set.

2) VQA-CP v2: VQA under changing priors (VQA-CP v2) is the new split of the data set VQA v2. Compared with VQA v2, VQA-CP v2 has changed the prior distributions of answers in train and test splits to avoid models, giving the most popular answer for the certain question type without understanding image content. The train set of VQA-CP v2 has 121k images, 438k questions, and 4.4M answers, while the test set has 98k images, 220k questions, and 2.2M answers, as shown in Table I.

3) FVQA: Fact-based VQA (FVQA) is designed to introduce supporting facts to help answer questions. Compared with VQA v2 and VQA-CP v2, FVQA contains the corresponding external knowledge, which was collected previously. However, FVQA includes fewer images and questions. The train set of FVQA has only 1100 images and 2927 corresponding questions, and the test set has 1090 images and 2899 corresponding questions, as shown in Table I.

B. Implementation Details

In accordance with the consistency in the baseline model, we adopt a multilayer structure similar to [12]. We set the number of layer G equal to 6, which has the best performance in the experiment. Also, we set the number of candidate answer words to 3129. The dimension d is 512 in multihead attention, and the number of head h is 8. Therefore, the dimension of each head is $d_h = d/h = 64$.

In the training stage, we use Adam [41] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The warm-up learning rate is from $2.5e^{-5}$ to $1e^{-4}$ in the first four epochs, and the learning rate keeps at $1e^{-4}$ from the fifth epoch to the tenth epoch. The learning rate decays by factor 0.2 in every two epochs until 13 epochs. The batch size fixes as 64 in the training and test stages. The dimension of visual feature is 2048 from Faster R-CNN, while words in questions and knowledge are encoded into a vector of dimension 300 with GloVe word embedding, and word vectors are processed through LSTM to get contextual vectors of dimension 512.

C. Detailed Image Feature

We take the finest VQA model MCANs as our experiment baseline model. The visual input to the baseline model is the image feature carrying ten to 100 object proposals extracted from Faster R-CNN. The textual input is the question feature processed through GloVe word embedding and LSTM. All our experiments are validated through comparison to the baseline model.

First, we set the number of object proposals for the image feature as 100 to provide richer details inside an input image. Table II shows the result of a different bounding box number. When the number of the object proposal equals 100, 67.77% on the validation set is provided, which is better compared with the baseline model with 10 to 100 object proposals. After the number of the object proposals is set to 50, the metric is reduced to 67.00%, which is lower than the baseline's 67.20%.

The experiment results in the first row of Table II represent our baseline model MCAN's performance, which covers 10 to 100 object proposals. The experiment results in the second row of Table II show the baseline model MCAN's performance with 100 object proposals. These two experiment results illustrate that more object proposals can improve the performance of our model by bringing richer visual information. The third row of Table II is the experiment result of the baseline model MCAN's performance with 50 object proposals. On the other hand, a lower number of object proposals mean that the model sees less and is possible to miss certain visual information, which might hurt the final performance.

TABLE II

ABLATION STUDY FOR NUMBER OF OBJECT PROPOSAL BOUNDING BOX IN BASELINE MODEL MCAN. ALL THE EXPERIMENT RESULTS ARE CONDUCTED ON THE VALIDATION SET

Experiment	Val accuracy
MCAN	67.20
box=100	**67.77**
box=50	67.00

D. External Knowledge

In Section III-B, we introduce two kinds of external information: 1) attribute and object information extracted from Faster R-CNN and 2) knowledge of object labels extracted from ConceptNet. To evaluate this external information, we present the ablation study on VQA v2 in Table III. When we employ the attribute and object information as our external information, the experiment result is 67.67% on the validation set. On the other hand, when we adopt the knowledge of object labels extracted from ConceptNet as our external information, the experiment result is 67.87%. Comparatively speaking, knowledge of object labels can bring more valuable information to infer the correct answer.

External knowledge of object labels is extracted from a large knowledge base ConceptNet. ConceptNet is a knowledge graph that connects entity words in natural language with labeled relations. ConceptNet has two types of relations, symmetric relations and asymmetric relations, and the total amount of relations is 40. However, not all the relations are suitable for compensating extra knowledge to the inference of the model. For example, a relation called "Antonym" represents the antonym of the word, and this helps little to the inference process. Thus, we refine these 40 relations to select ones that are capable to describe the characters, abilities of the entity word, or relations linking two entity words.

The relations that we used in this article are divided into three types: properties of entity words (e.g., HasProperty, DefinedAs, IsA, HasA, and HasContext), the spatial location of objects (e.g., AtLocation, LocatedNear, PartOf, and SymbolOf), and the tendency of objects (e.g., MadeOf, UsedFor, ReceivesAction, RelatedTo, CapableOf, and MannerOf).

When we do not introduce any knowledge from knowledge graph ConceptNet, the experiment result of the baseline model is 67.20% in the validation set, as shown in the first row in Table IV. Moreover, through experimental comparisons, we find that if the extracted number of explicit object labels and implicit object labels K is set to 5, the performance is the best 67.87%, as seen in the third row of Table IV.

TABLE III

ABLATION STUDY FOR TWO KINDS OF EXTERNAL INFORMATION. COMPARED WITH ATTRIBUTE AND OBJECT INFORMATION EXTRACTED FROM FASTER R-CNN, KNOWLEDGE OF OBJECT LABELS EXTRACTED FROM CONCEPTNET OUTPERFORMS IN THE ABLATION STUDY

Experiment	Val accuracy
attribute and object	67.67
knowledge of object labels	**67.87**

TABLE IV

ABLATION STUDY FOR THE NUMBER OF EXTERNAL KNOWLEDGE FROM THE EXTERNAL KNOWLEDGE BASE. MCAN IN THE FIRST ROW INDICATES THE BASELINE MODEL WITHOUT INTRODUCING ANY KNOWLEDGE. OTHER ROWS INDICATE THE PERFORMANCE OF THE BASELINE MODEL WITH A CERTAIN NUMBER OF OBJECTS AND RELATIONS

#Knowledge	#Relations	Val accuracy
K=0	R=0	67.20
K=10	R=2	67.72
K=5	R=2	**67.87**
K=3	R=2	67.82
K=5	R=5	67.78

When introducing more object labels by setting K equal to 10, the accuracy of the experiment decreases in the second row. Although more detailed information is brought in, objects related to the question are not as many as expected. More labels include more disturbing information; thus, a negative influence is added to the inference of our model. If the number of object labels is set to 3, the experiment result drops to 67.82% because less information is exposed to the model, which means the model can learn less common sense from the knowledge graph. By contrast, the model might give plausible answers when the model learns inadequately.

On the other hand, when the number of relations for each object label R equals 2, we can get the best performance 67.87% in the experiment. If we introduce more relations and set R equal to 5, the accuracy of the experiment instead decreases, as shown in the fifth row in Table IV. Each knowledge related to one entity word in ConceptNet has its weight, which indicates the degree of the corresponding knowledge that we can rely on. The knowledge weight of entity words decreases when the number of knowledge increases. Some knowledge with low scores does not help during the inference process when introducing more knowledge; thus, the adverse effect is produced instead.

E. Adaptive Score Attention Module

Further improvement can be obtained by sufficiently utilizing attended knowledge features $\mathbf{F}^G$ and attended visual feature $\mathbf{X}^G$ provided by the symmetric modulized network. However, simply combining them provides constrained advantages in the experiment. Three frequently used combination methods between attended knowledge feature $\mathbf{F}^G$ and attended visual feature $\mathbf{X}^G$ are employed in our experiments: concatenation, summation, and wise-product. The experimental results are shown in Table V, where three easy combination methods improve little compared with the method that only uses attended visual feature $\mathbf{X}^G$ and provides the accuracy of 67.77% shown in the second row. The third row in Table V shows that the attended knowledge feature $\mathbf{F}^G$ and the attended visual feature $\mathbf{X}^G$ combine in a concatenation way. The experiment accuracy is 67.67% on average after three experiments, which is much lower than the baseline. The experiment result of the summation of two kinds of attended feature is 67.77%, as shown in the fourth row. At the same time, the experiment result of wise-product between two kinds of attended feature gets 67.76% in the fifth row, near the summation result. All three simply combined methods fail to get better performance than the method only with attended visual feature $\mathbf{X}^G$.

TABLE V

ABLATION STUDY FOR METHODS OF MERGING ATTENDED IMAGE FEATURE $\mathbf{X}^G$ AND ATTENDED FACT FEATURE $\mathbf{F}^G$. MCAN IS THE BASELINE MODEL WITHOUT ANY KNOWLEDGE FEATURE. ATTENDED VISUAL FEATURE $\mathbf{X}^G$ INDICATES BASELINE ONLY WITH VISUAL FEATURES. CONCATENATION, SUMMATION, AND WISE-PRODUCT REPRESENT THREE FREQUENTLY USED METHODS. ADAPTIVE SCORE ATTENTION IS THE EFFECTIVE METHOD THAT WE PROPOSED FOR BOTH ATTENDED IMAGE FEATURE AND ATTENDED KNOWLEDGE FEATURE

Experiment	Val accuracy
MCAN	67.20
attended visual feature $\mathbf{X}^G$	67.77
concatenation	67.67
summation	67.77
wise-product	67.76
adaptive score attention	**67.87**

After attended knowledge feature $\mathbf{F}^G$ and attended visual feature $\mathbf{X}^G$ are combined through the newly designed adaptive score attention module, the accuracy improves to 67.87%, which is a relatively large margin. From the experiments in the ablation study, the adaptive score attention module can combine image feature and knowledge feature effectively, keep a balance between these two kinds of features, and, thus, fully enhance the performance of our model in the experiment. Concretely, the adaptive score attention module can intelligently learn the importance between image and knowledge on certain questions. When the focus point of the question is related to the image instead of the knowledge, the importance of image is much larger than the knowledge. Therefore, after processed by the adaptive score attention module, the attended visual feature $\mathbf{X}^G$ gets the score close to 1, and the attended knowledge feature gets the score near to 0. In this manner, our model can get expected performance in utilizing the adaptive score attention module to automatically select the most related source feature about the question.

F. Qualitative Analysis

As aforementioned, when answering questions, only relying on the visual information is not adequate. If we require the model to be as intelligent as a human being, we should leverage the external knowledge base (e.g., ConceptNet) to let our model receive common sense or knowledge that the model lacks to compensate for the inadequacy of visual cues. For example, for the image in the top left row of Fig. 5, a model should answer the question "Why are some people

Q: Why are some people holding umbrellas?

K: ['An umbrella is for protection from the **rain**',
'Something you find in a closet is an umbrella',
'An umbrella is a device to protect something',
'Branch is related to tree',
'Squirrel is related to tree']

GT： raining
MCAN: shade
KAN(Ours): raining

Q: Why would someone want an oven like that?

K: ['Something you find in the oven is food',
'An oven is used for **cooking**',
'Something you find in the oven is racks',
'Something you find in the kitchen is food',
'A kitchen is used for **cooking** food']

GT: cooking
MCAN: yes
KAN(Ours): cooking

Q: Why are these giraffes indoors?

K: ['Something you find at a **zoo** is giraffe',
'Okapi is a type of giraffe',
'A giraffe can be male',
'Door is related to entrance',
'Door is related to opening']

GT: zoo
MCAN: yes
KAN(Ours): zoo

Q: Is this a lake?

K: ['You are likely to find water in a lake',
'Lake is related to water',
'Nest is related to bird',
'You are likely to find a fish in water.',
'River is related to water']

GT: yes
MCAN: yes
KAN(Ours): yes

Q: How many beds?

K: ['Bed is related to sleeping',
'Bed is related to sleep',
'Bed is related to furniture',
'Chair is related to sitting',
'Seat is related to chair']

GT: 1
MCAN: 1
KAB(Ours): 1

Q: What are these people sitting on?

K: ['Crowd is related to people',
'Boy is related to man',
'Man is related to person',
'Window is related to glass',
'Glass is related to window']

GT: grass
MCAN: grass
KAB(Ours): grass

Fig. 5. Qualitative analysis of different kinds of questions. Image–question pairs of the top row in the figure highlight the impact of knowledge as images do not have enough visual information to answer the question. For instance, in the top left row of the figure, the question (Q) focuses on the reason why some people hold the umbrellas; thus, simple image content cannot provide enough information to give the correct answer (A). However, under the guidance of external knowledge (K), our model can seek the answer successfully. Our answers in green are correct but the MCAN model's answers in red are incorrect. On the other hand, image–question pairs of the bottom row in the figure explicate the importance of visual features. When the external knowledge does not help to answer the questions, our model and MCAN model can both provide correct answers.

holding umbrellas?" and the true answer is "rain." Only object "umbrella" shows in the image without further information about "rain" so previous methods that rely on knowledge only from the image face difficulty to provide an accurate solution.

However, introducing external knowledge facilitates our model to provide an accurate answer. Concretely, we query all relevant knowledge about the "umbrella" from ConceptNet and order them by the degree of their corresponding credibility. The first knowledge of "umbrella" is "An umbrella is for protection from the rain," which is beneficial for our model to provide the correct answer. Now, we have both images that cannot directly provide the right solution and knowledge that teaches the model what a fitting answer can be, and if we utilize these two source information without distinguishing, our model instead decreases the performance. The freshly proposed adaptive score attention module can adaptively determine more relevant information to support our model locate true solution "rain" from external knowledge and image.

On the other hand, when a question focuses on visual objects, rich and detailed visual information lets our model clearly "see" the whole image, thus answers the question accurately. For example, the question, "How many beds?"

TABLE VI

ACCURACY RESULT ON THE TEST-DEV SET AND TEST-STD SET WITH THE STATE-OF-THE-ART METHODS ON THE VQA V2 DATA SET. OUR BEST SINGLE MODEL IS TRAINED ON THE TRAIN+VAL SPLITS PLUS THE AUGMENTED DATA SET VG, WHICH IS THE SUBSET OF VISUAL GENOME

Model	Test-dev				Test-std
	All	Y/N	Num	Other	All
Bottom-Up	65.32	81.82	44.21	56.05	65.67
BAN	69.52	85.31	50.93	60.26	-
BAN+Counter	70.04	85.42	54.04	60.52	70.35
DFAF	70.22	86.09	53.32	60.49	70.34
MCAN	70.63	86.82	53.26	60.72	70.90
Ours	**71.56**	**87.69**	**54.51**	**61.64**	**71.84**

TABLE VII

RESULTS ON THE VQA-CP V2 [39] DATA SET. OUR MODEL OUTPERFORMS OTHER MODELS ON NUMBER, OTHER, AND OVERALL TASKS

Model	Yes/No	Num	Other	Overall
HAN	52.25	13.79	20.33	28.65
GVQA	**57.99**	13.68	22.14	31.30
MuRel	42.85	13.17	45.04	39.54
Ours	42.12	**15.52**	**50.28**	**42.60**

concentrates at as many beds as possible within an image. By extracting more valuable visual details through Faster R-CNN, our model is more capable of fulfilling the question, which focuses on counting on visual objects. However, the object label "beds" from the question does not help to the model. The object label "beds" knowledge with most weight scores provided from the external knowledge base ConceptNet, such as "Bed is related to sleeping," promotes little function for elucidating the question. Therefore, our adaptive score attention module automatically filters the external knowledge and selects the detailed information from the input image to present a fitting answer.

G. Comparison With Other Methods

1) VQA v2: Through a series of ablation studies, we select the model with the adaptive score attention module that provides the best result to compare with other existing methods. We score 71.51% of overall accuracy on the Test-dev data set and perform 0.88% better than the baseline model, as shown in Table VI. Also, our model obtains 71.84% of the overall accuracy on the Test-std data set. At the same time, compared with the previous models, our model also provides the best results on the Yes/No, Number, and Other tasks, which indicates that our model is designed in a more advanced way. Although BAN+Counter model [44] used an additional module "Counter" [45] for number task, our model still achieves a better score on the number task, which indicates that our model has comprehensive inference capability.

2) VQA-CP v2: VQA-CP v2 is a new data set to overcome the bias of training data of the original VQA v2 data set. To indicate the generalization ability of our model, we take extra experiments on the data set VQA-CP v2. Our model obtains the best result of 42.60% on the overall task. Moreover, our model outperforms other models on two subtasks of Number and Other. The GVQA model is designed for Yes/No subtask and has the best performance on it. In general, our model has outstanding performance on the overall task and balanced scores on three subtasks.

TABLE VIII

TOP-ONE OVERALL ACCURACY RESULT WITH OTHER METHODS ON THE FVQA DATA SET. OUR BEST SINGLE MODEL OUTPERFORMS THE BEST RESULT REPORTED IN FVQA IN TOP-ONE ACCURACY

Model	Overall Acc. $\pm$ Std (%)
	Top-1
SVM-Question	10.37 ± 0.80
SVM-Image	18.41 ± 1.07
Hie-Question+Image	33.70 ± 1.18
Hie-Question+Image+Pre-VQA	43.14 ± 0.61
FVQA	63.63 ± 0.73
Ours	$\mathbf{66.39 \pm 0.50}$

3) FVQA: FVQA is collected as a knowledge-based VQA data set. To indicate the capacity in answering knowledge-based visual questions, we employ our model to conduct an additional experiment on the data set FVQA. Our model obtains $66.39 \pm 0.50\%$ top-one overall accuracy in five random splits of the data set, as shown in Table VIII. Our experiment's overall accuracy outperforms the best top-one overall accuracy reported in FVQA, which is $63.63 \pm 0.73\%$.

V. CONCLUSION

In this work, we have proposed a novel KAN for VQA, which introduced richer visual information and compensated common sense from an external knowledge base. Furthermore, for different types of questions, we have made our model adaptively balance the importance between visual information and external knowledge. Thus, we have introduced a new adaptive score attention module that automatically chooses a suitable information source depending on the type of question. Experimental results have shown that our model possessed the state-of-the-art outcomes on the VQA v2, VQA-CP v2, and FVQA.

REFERENCES

[1] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[2] 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," 2016, *arXiv:1606.01847*. [Online]. Available: http://arxiv.org/abs/1606.01847

[3] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[4] X. Li *et al.*, "Beyond RNNs: Positional self-attention with co-attention for video question answering," in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 8, 2019, pp. 8658–8665.

[5] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[10] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[11] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1045–1048.

[12] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6087–6096.

[13] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: http://arxiv.org/abs/1812.08434

[14] M. Kinderkhedia, "Learning representations of graph data-a survey," 2019, *arXiv:1906.02989*. [Online]. Available: http://arxiv.org/abs/1906.02989

[15] A. Singh *et al.*, "Towards VQA models that can read," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8317–8326.

[16] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–10.

[17] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[18] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.

[19] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2296–2304.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4995–5004.

[22] X. Li *et al.*, "Learnable aggregating net with diversity learning for video question answering," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1166–1174.

[23] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.

[24] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, May 2020.

[25] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6281–6290.

[28] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4584–4593.

[29] X. Pan *et al.*, "Improving question answering with external knowledge," 2019, *arXiv:1902.00993*. [Online]. Available: http://arxiv.org/abs/1902.00993

[30] S. Park, S. Kwon, B. Kim, S. Han, H. Shim, and G. G. Lee, "Question answering system using multiple information source and open type answer merge," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, 2015, pp. 111–115.

[31] Z. Dai, L. Li, and W. Xu, "CFO: Conditional focused neural question answering with large-scale knowledge bases," 2016, *arXiv:1606.01994*. [Online]. Available: http://arxiv.org/abs/1606.01994

[32] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," 2015, *arXiv:1511.02570*. [Online]. Available: http://arxiv.org/abs/1511.02570

[33] J. Hoffart *et al.*, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Commun. ACM*, vol. 52, no. 4, pp. 56–64, 2009.

[34] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a Web of open data," in *The Semantic Web*. Berlin, Germany: Springer, 2007, pp. 722–735.

[35] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1969–1978.

[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[38] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.

[39] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4971–4980.

[40] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "FVQA: Fact-based visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2413–2427, Oct. 2018.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, "Unified binary generative adversarial network for image retrieval and compression," *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2243–2264, Sep. 2020.

[43] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[44] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.

[45] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," 2018, *arXiv:1802.05766*. [Online]. Available: http://arxiv.org/abs/1802.05766

[46] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.

Liyang Zhang received the B.E. and M.S. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2017 and 2020, respectively.

His research interests include computer vision and natural language processing.

Shuaicheng Liu (Member, IEEE) received the B.E. degree from Sichuan University, Chengdu, China, in 2008, and the M.S. and Ph.D. degrees from the National University of Singapore (NUS), Singapore, in 2010 and 2014, respectively.

In 2014, he joined the University of Electronic Science and Technology of China (UESTC), Chengdu, where he is currently an Associate Professor with the School of Information and Communication Engineering. His research interests include computer vision and computer graphics.

Donghao Liu received the M.Sc. degree from The University of Edinburgh, Edinburgh, U.K., in 2016.

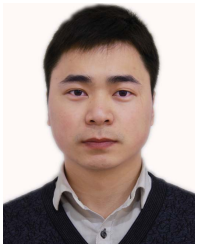
He is currently an Assistant Deep Learning Research Engineer with Megvii, Chengdu, China. He is also working on computational photography pipelines enhanced by deep neural networks.

Pengpeng Zeng is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China.

He is also conducting research on visual understanding.

Xiangpeng Li is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China.

He is also conducting research on visual understanding.



Jingkuan Song (Senior Member, IEEE) is currently a Professor with the University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests include large-scale multimedia retrieval, image/video segmentation and image/video understanding using hashing, graph learning, and deep learning techniques.

Dr. Song has been an AC/SPC/PC Member of IEEE Conference on Computer Vision and Pattern Recognition for the term 2018–2021, and so on. He was the winner of the Best Paper Award in International Conference on Pattern Recognition, Mexico, in 2016, the Best Student Paper Award in Australian Database Conference, Australia, in 2017, and the Best Paper Honorable Mention Award, Japan, in 2017.

Lianli Gao (Member, IEEE) received the Ph.D. degree in information technology from The University of Queensland (UQ), Brisbane, QLD, Australia, in 2015.

She is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. Especially, she is mainly focusing on integrating natural language for visual content understanding.

Dr. Gao was the winner of the IEEE TRANSACTIONS ON MULTIMEDIA 2020 Prize Paper Award, the Best Student Paper Award in the Australian Database Conference, Australia, in 2017, the IEEE TCMC Rising Star Award in 2020, and the ALIBABA Academic Young Fellow.