MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering

Yang Ding[1,2], **Jing Yu**[1,2*], **Bang Liu**[3,4†], **Yue Hu**[1,2], **Mingxin Cui**[1,2], **Qi Wu**[5]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Université de Montréal, Canada [4]Mila - Quebec AI Institute, Canada
[5]University of Adelaide, Australia
{dingyang, yujing02, huyue, cuimingxin}@iie.ac.cn, bang.liu@umontreal.ca
qi.wu01@adelaide.edu.au

Abstract

Knowledge-based visual question answering requires the ability of associating external knowledge for open-ended cross-modal scene understanding. One limitation of existing solutions is that they capture relevant knowledge from text-only knowledge bases, which merely contain facts expressed by first-order predicates or language descriptions while lacking complex but indispensable multimodal knowledge for visual understanding. How to construct vision-relevant and explainable multimodal knowledge for the VQA scenario has been less studied. In this paper, we propose MuKEA to represent multimodal knowledge by an explicit triplet to correlate visual objects and fact answers with implicit relations. To bridge the heterogeneous gap, we propose three objective losses to learn the triplet representations from complementary views: embedding structure, topological relation and semantic space. By adopting a pre-training and fine-tuning learning strategy, both basic and domain-specific multimodal knowledge are progressively accumulated for answer prediction. We outperform the state-of-the-art by 3.35% and 6.08% respectively on two challenging knowledge-required datasets: OK-VQA and KRVQA. Experimental results prove the complementary benefits of the multimodal knowledge with existing knowledge bases and the advantages of our end-to-end framework over the existing pipeline methods. The code is available at https://github.com/AndersonStra/MuKEA.

1. Introduction

Visual Question Answering based on external Knowledge Bases (KB-VQA) [37] requires an AI agent to answer a question by incorporating knowledge about the world beyond what the question and the image contains. Despite the great success in VQA tasks [11, 40], KB-VQA is more challenging for models to achieve human-like ability of open-ended cross-modal scene understanding associating with external knowledge. Therefore, how to appropriately represent and leverage knowledge in such cross-modal scenario becomes a core problem of KB-VQA.

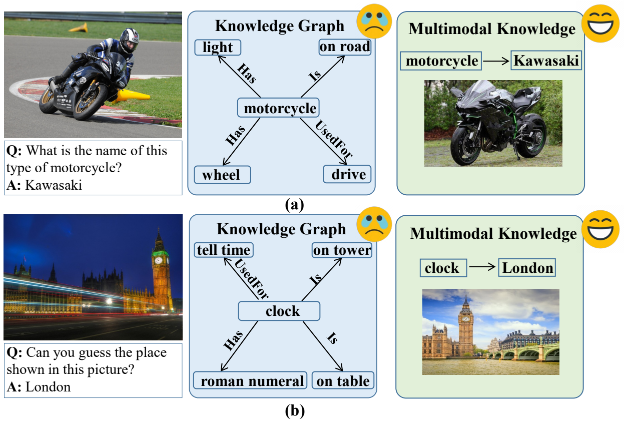


Figure 1. An illustration of our motivation. Compared with rigid facts in the knowledge graph, multimodal knowledge for depicting complex and inexpressible facts is indispensable in both open-ended object understanding (a) and scene understanding (b).

Most of recent works [9, 23, 46] focus on capturing relevant knowledge from structured knowledge graphs, such as ConceptNet [18] and DBpedia [4], or unstructured/semi-structured knowledge, like Wikipedia [1] and Visual Genome [15]. Though these knowledge bases provide high-quality knowledge by large-scale human annotations, the information is generally limited to the definite facts that can be explicitly expressed by natural language or simple triplets with first-order predicate. Therefore, such knowledge bases are quite difficult to represent high-order predi-

[*] Corresponding author.
[†] Canada CIFAR AI Chair.

cate and multimodal knowledge, which is essential for human to tackle complex problems. Considering the question in Figure 1(a), the agent needs visual knowledge of motorcycle appearance in each brand to identify the given motorcycle, but the knowledge graph lacks of such instantiated information. Besides object understanding, implicit visual knowledge in mind mostly dominate over the rigid facts when humans are asked for simple scene discrimination like the question 'Can you guess the place?' in Figure 1(b). *How to represent and accumulate the complex multimodal knowledge in the VQA scenario while maintaining the advantages of traditional knowledge graph in explainable reasoning is an essential but less studied problem.*

Current progress [17, 30, 33] in emerging multimodal knowledge graph aims to correlate visual content with textual facts to form the augmented knowledge graph. The typical solutions can be divided into two categories: parsing images and texts to structured representations and grounding event/entities across modalities [13, 17, 39], or simply aligning the entities in existing knowledge graphs with related images [30, 33]. However, such multimodal knowledge graphs in essence still represent knowledge via the first-order predicate, which fails to model the high-order complex relationships such as the relationship between 'clock' and 'London' in Figure 1(b).

In this paper, we propose a novel ***Mu****ltimodal* ***K****nowledge* ***E****xtraction and* ***A****ccumulation* framework (MuKEA) for KB-VQA task. Independent of existing knowledge bases, the core mechanism behind MuKEA is to accumulate multimodal knowledge with complex relationships from observation of VQA samples, and perform explainable reasoning based on the self-accumulated knowledge. To this end, we first propose a novel schema to represent multimodal knowledge unit by an explicit triplet, where the visual objects referred by the question are embedded in the head entity, the embedding of the fact answer is kept in the tail entity, and the implicit relation between the head and the tail is expressed by the relation. We propose three objective loss functions to learn the representations of the triplets from coarse to fine by contrasting positive and negative triplets, aligning ground-truth triplets, and refining entity representations. A pre-training and fine-tuning learning strategy is then proposed to progressively accumulate multimodal knowledge from both out-domain and in-domain VQA samples for explainable reasoning.

The main contributions of this work are as follows:

(1) We propose an end-to-end multimodal knowledge representation learning framework, which first models the inexpressible multimodal facts by explicit triplets and provides complementary knowledge with the existing knowledge graphs and unstructured knowledge bases.

(2) We exploit a pre-training and fine-tuning strategy to accumulate both out-domain and in-domain knowledge to form a neural multimodal knowledge base. It supports automatic knowledge association and answer prediction, which gets rid of the cascading error in existing 'knowledge retrieve and read' pipeline [23, 46].

(3) Our model with strong generalization ability outperforms the state-of-the-art models by 3.35% and 6.08% respectively on two challenging KB-VQA datasets: OK-VQA [24] and KRVQA [7]. The good performance can be well explained by visualizing the relevant multimodal knowledge triplets explicitly.

2. Related Work

Knowledge-based Visual Question Answering. Most of recent works are based on 'knowledge retrieve and read' pipeline, which requires highly-relevant knowledge to support knowledge reasoning. Structured knowledge based methods like [9] is based on ConceptNet [18] to introduce knowledge in the form of triplet with first-order predicate. Unstructured knowledge based methods [24] retrieve knowledge from Wikipedia [1] and encode relevant text in a memory network for further reasoning. However, knowledge described in nature language lacks visual information to assist cross-modal understanding. For the above challenge, [44] augments the knowledge graph YAGO [31] with related images to serve as multimodal knowledge. However, such graph in essence still represents knowledge via the first-order predicate. To go one step further, we extract multimodal information to represent high-order complex relations and represent multimodal knowledge by explicit triplets for explainable reasoning.

From the view of model framework, most of recent works are based on the 'retrieve and read' pipeline, which first retrieve the relevant facts from knowledge bases and then perform explicit reasoning on the knowledge graph [25, 36, 37], or fusing the implicit knowledge embedding with the corresponding image and questions for answer classification [9, 16, 24]. All of these methods rely on object labels to retrieve external knowledge, which inevitably introduces irrelevant knowledge and leads to cascading error. There are also end-to-end methods based on implicit knowledge like pre-trained models [20, 23, 34]. However, such implicit knowledge mainly captures the co-occurrence of image-question-answer triplet instead of explainable and refined knowledge. In this paper, we propose an end-to-end multimodal knowledge extraction and accumulation framework with interpretable triplet knowledge.

Multimodal Knowledge Graph. The emerging multimodal knowledge graph works [17, 30] aim to correlate visual content with textual facts to form the augmented knowledge graph. One typical solution parses images and texts to structured representations first and grounds event/entities across modalities. The key problem lies in intra-modal relation extraction and cross-modal entity link-

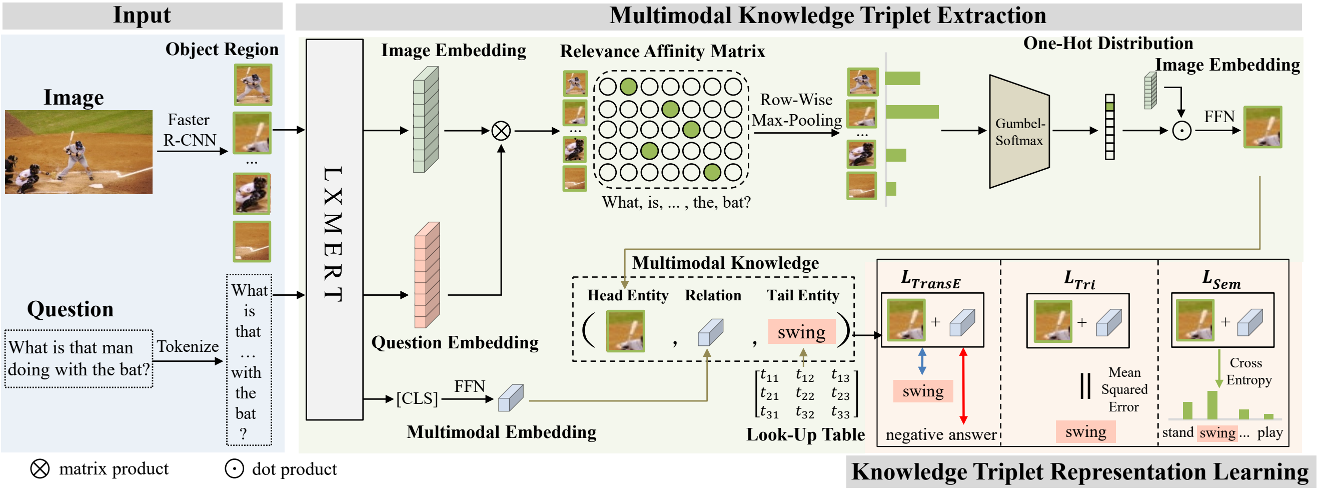


Figure 2. An overview of our model. The model contains two modules: Multimodal Knowledge Triplet Extraction aims to extract multimodal knowledge triplets from samples and Knowledge Triplet Representation Learning aims to unifiedly learn the triplet representation.

ing. Specifically, [17, 26] learn knowledge from structured textual and visual data and maintain the triplet structure for entity alignment. [13] utilizes RDF [22] knowledge graphs to represent multimodal information based on graph alignment and lacks multimodal correlation. Another kind of solutions directly links the entities in existing knowledge graphs with relevant images. [30] adds images to expand the entity representation in YAGO [31]. However, all of these methods in essence still represent knowledge via the first-order predicate described by nature language, which fail to model the high-order complex relationships.

3. Methodology

Given an image I and a question Q, the KB-VQA task aims to predict an answer A supported by external knowledge beyond the given visual and textual content. We accumulate triplet-formed multimodal knowledge to serve as the external knowledge and directly infer the answer in an end-to-end mode. Figure 2 gives detailed illustration of our model. We first introduce a novel schema of extracting multimodal knowledge triplets from unstructured image-question-answer samples based on the pre-trained vision-language model. Then we propose three objective losses to learn the triplet embeddings that accurately depict question-attended visual content (head embeddings), question-desired fact answer (tail embeddings), and the implicit relation between the two (relation embeddings). By training with both out-domain and in-domain data, our model accumulates a wide range of multimodal knowledge and associates the optimal fact for answer prediction.

3.1. Multimodal Knowledge Triplet Extraction

In the VQA scenario, we define the complex and inexpressible facts as multimodal knowledge in the form of triplet, *i.e.* (h, r, t), where h contains visual content in the image focused by the question, t is a representation of the answer given the question-image pair, and r depicts the implicit relationship between h and t containing multimodal information. The triplet construction process mainly consists of the following four parts:

Image and Question Encoding. Since the pre-trained vision-language models are strong at modeling the intra-modal and cross-modal implicit correlations, we first utilize the pre-trained model LXMERT [34] to encode the question and image for further multimodal knowledge triplet extraction. We apply Faster R-CNN [32] to detect a set of objects $\mathcal{O} = \{o_i\}_{i=1}^{K}$ (K = 36) in I and represent each object o_i by a visual feature vector $\boldsymbol{f}_i \in \mathbb{R}^{d_f}$ (d_f = 2048) and a spatial feature vector $\boldsymbol{b}_i \in \mathbb{R}^{d_b}(d_b = 4)$ as in [46]. We tokenize a question Q using WordPiece [38] and obtain a sequence of D tokens. We feed the visual features $\{\boldsymbol{f}_i\}_{i=1}^{K}$ and $\{\boldsymbol{b}_i\}_{i=1}^{K}$, and question tokens into the pre-trained LXMERT, obtaining the visual embeddings of $\mathcal{O}$ denoted as $\boldsymbol{V} \in \mathbb{R}^{K\times d_v}$ (d_v = 768) and the token embeddings denoted as $\boldsymbol{Q} \in \mathbb{R}^{D\times d_v}$.

Head Entity Extraction. We define the head entity as the visual object and its context in the image that is most relevant to the question. To this end, we firstly evaluate the relevance of each object in the image to each token in the question by computing the question-guided object-question relevance affinity matrix $\boldsymbol{A}$ as:

$$\boldsymbol{A} = (\mathbf{W}_1\boldsymbol{Q})^T(\mathbf{W}_2\boldsymbol{V}) \tag{1}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learned parameters.

Under the guidance of the relevance affinity matrix, we then select one object in $\mathcal{O}$ as the most relevant visual content to the question. Since LXMERT models the implicit correlations among all the objects, it is noteworthy that the selected question-centric object already contains its contextual information, which provides indispensable clues for answering questions that involve multiple objects. Specifically, we compute the row-wise max-pooling on $\boldsymbol{A}$ to evaluate the relevance of each object o_i to the question as:

$$\boldsymbol{a}_i^{v-q} = \max_j \boldsymbol{A}_{i,j} \tag{2}$$

Then hard attention instead soft attention is applied to select the most relevant object as the head entity based on $\{\boldsymbol{a}_i^{v-q}\}_{i=1}^{K}$. Compared with soft attention, hard attention provides more stable and explainable visual content for multimodal knowledge representation, which is also easier for combining with exiting knowledge graph by entity linking. Here we conduct Gumbel-Softmax [12] to obtain the approximate one-hot categorical distribution. The attention weight for object o_i is computed as:

$$\alpha_i = \frac{\exp((\log(\boldsymbol{a}_i^{v-q}) + g_i)/\tau)}{\sum_{j=1}^{K}\exp((\log(\boldsymbol{a}_j^{v-q}) + g_j)/\tau)} \tag{3}$$

where $\{g_i\}_{i=1}^{K}$ are *i.i.d.* samples drawn from Gumbel(0,1)[1], and τ is the softmax temperature. Finally, we gather the question-centric object information and obtain the head entity representation $\boldsymbol{h}$ as:

$$\boldsymbol{h} = \mathrm{FFN}(\sum_{i=1}^{K}\alpha_i \boldsymbol{v}_i) \tag{4}$$

where $\boldsymbol{v}_i \in \boldsymbol{V}$ and FFN denotes a feed-forward network that contains two fully connected layers.

Relation Extraction. Different from the relation in traditional knowledge graph that depicts the first-order predicate independent of specific visual scenario, we define the relation in multimodal knowledge as the complex implicit relation between the observed instantiated object and the corresponding fact answer. Since LXMERT captures the implicit correlations between the image and the question via the self-attention mechanism in the hierarchical transformers, we extract the cross-modal representation from the [CLS] token, and feed it into a FFN layer to obtain the relation embedding, which is denoted as $\boldsymbol{r}$.

Tail Entity Extraction. We define the tail entity as the answer in an image-question-answer sample, which reveals a specific aspect of facts regarding to the visual object referred by the question. In the training stage, we set ground-truth answer as the tail entity to learn its representation $\boldsymbol{t}$ from scratch (details in Section 3.2). In the inference stage, we define the KB-VQA task as a multimodal knowledge graph completion problem and globally assess the knowledge in our neural multimodal knowledge base to predict the optimal tail entity as the answer (details in Section 3.3).

3.2. Knowledge Triplet Representation Learning

Since each component within a triplet contains modality-different and semantic-specific information, we propose three loss functions to unifiedly learn the triplet representation in order to bridge the heterogeneous gap as well as semantic gap. The three losses constrain the triplet representation from complementary views: the *Triplet TransE Loss* preserves the embedding structure by contrasting positive and negative triplets. The *Triplet Consistency Loss* further forces the three embeddings within a triplet to keep the strict topological relation, and the *Semantic Consistency Loss* maps the embeddings into a common semantic space for direct comparison among multimodal content.

Triplet TransE Loss. Inspired by the knowledge embedding method TransE [6] in the traditional knowledge graph field, we apply TransE-like objective loss as a structure-preserving constraint in our multimodal scenario. Given an image-question pair, let $\mathcal{A}^+$ and $\mathcal{A}^-$ denote the sets of correct (positive) and incorrect (negative) answers, respectively. Let $\boldsymbol{h}$ and $\boldsymbol{r}$ denote the corresponding extracted head and tail entity representations. We push the distance between $\boldsymbol{h}+\boldsymbol{r}$ and each positive tail $\boldsymbol{t}^+ \in \mathcal{A}^+$ to be smaller than the distance between $\boldsymbol{h}+\boldsymbol{r}$ and each negative tail $\boldsymbol{t}^- \in \mathcal{A}^-$ by a certain margin γ:

$$\mathcal{L}_{\mathrm{TransE}} = \sum_{\boldsymbol{t}^+\in\mathcal{A}^+}\sum_{\boldsymbol{t}^-\in\mathcal{A}^-}[\gamma + \mathrm{d}(\boldsymbol{h}+\boldsymbol{r},\boldsymbol{t}^+) - \mathrm{d}(\boldsymbol{h}+\boldsymbol{r},\boldsymbol{t}^-)]_+ \tag{5}$$

where $[\cdot]_+ \triangleq \max(0,\cdot)$ and $\mathrm{d}(\cdot,\cdot)$ denotes the cosine distance following the settings in [21].

Triplet Consistency Loss. The issue of the above TransE loss is that once the distance between the positive pairs is smaller than the negative pairs by margin γ during training, the model will stop learning from the triplet. To further push the embeddings to satisfy the strict topological relation, we apply Mean Squared Error (MSE) criterion to constrain the representations on top of each positive triplet as:

$$\mathcal{L}_{\mathrm{Tri}} = \mathrm{MSE}(\boldsymbol{h}+\boldsymbol{r},\boldsymbol{t}^+) \tag{6}$$

Semantic Consistency Loss. We randomly initialize a look-up table of tail entities and learn their representations jointly with the head and the relation. Each tail entity in the look-up table $\boldsymbol{T}$ corresponds to an unique answer in the training VQA samples. To introduce the semantics of answer in tail representation while narrowing the heterogeneous gap between text-formed tail entity and multimodal-formed head entity and relation, we classify the triplet over the tail vocabulary and force the model to select the ground-truth tail (answer) by the negative log likelihood loss:

$$P(\boldsymbol{t}^+) = \mathrm{softmax}((\boldsymbol{T})^T(\boldsymbol{h}+\boldsymbol{r})) \tag{7}$$

$$\mathcal{L}_{\mathrm{Sem}} = -\log(P(\boldsymbol{t}^+)) \tag{8}$$

[1] The Gumbel (0,1) distribution can be sampled using inverse transform sampling by drawing $u \sim$ Uniform(0,1) and computing $g = -\log(-\log(u)))$

Method	Knowledge Resources	Accuracy
ArticleNet (AN) [24]	Wikipedia	5.28
Q-only [24]	—	14.93
BAN [14]	—	25.17
+AN [24]	Wikipedia	25.61
+ KG-AUG [16]	Wikipedia + ConceptNet	26.71
MUTAN [5]	—	26.41
+ AN [24]	Wikipedia	27.84
Mucko [46]	ConceptNet	29.20
GRUC [41]	ConceptNet	29.87
KM^4 [44]	multimodal knowledge from OK-VQA	31.32
ViLBERT [20]	—	31.35
LXMERT [34]	—	32.04
KRISP(w/o mm pre.) [23]	DBpedia + ConceptNet + VisualGenome + haspartKB	32.31
KRISP(w/ mm pre.) [23]	DBpedia + ConceptNet + VisualGenome + haspartKB	38.90
ConceptBert [9]	ConceptNet	33.66
Knowledge is Power [45]	YAGO3	39.24
MuKEA	multimodal knowledge from VQA 2.0 and OK-VQA	**42.59**

Table 1. State-of-the-art comparison on OK-VQA dataset. The middle column lists the external knowledge sources, if any, used in each VQA system. The rows in the middle part list the method based on pre-trained model.

where $P(\boldsymbol{t}^+)$ is the predicted probability of ground-truth tail $\boldsymbol{t}^+$. In summary, our final loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{TransE}} + \mathcal{L}_{\text{Tri}} + \mathcal{L}_{\text{Sem}} \tag{9}$$

3.3. Knowledge Accumulation and Prediction

We adopt a two-stage training strategy to accumulate multimodal knowledge progressively: (1) pre-training on the VQA 2.0 dataset [10] to accumulate basic visual-dominant knowledge and then (2) fine-tuning on the training data of downstream KB-VQA task to accumulate more complex domain-specific multimodal knowledge. All questions in VQA 2.0 are divided into three categories: *Yes/No*, *Number*, and *Other*. Since answers in the first two categories can not serve as the fact knowledge, we only keep *Other* type questions for pre-training.

In the inference stage, we regard the answer prediction as a multimodal knowledge graph completion problem. Given an image and a question, we feed them into the network and obtain the embeddings of the head entity $\boldsymbol{h}_{inf}$ and the relation $\boldsymbol{r}_{inf}$. We compute the distance between $\boldsymbol{h}_{inf}+\boldsymbol{r}_{inf}$ and each tail entity $\boldsymbol{t}_i$ in the look-up table $\boldsymbol{T}$, and select the tail entity with the minimum distance as:

$$\boldsymbol{t}_{inf} = \mathop{\arg\min}_{\boldsymbol{t}_i \in \boldsymbol{T}} \mathrm{d}(\boldsymbol{h}_{inf} + \boldsymbol{r}_{inf}, \boldsymbol{t}_i) \tag{10}$$

The answer corresponding to the optimal tail entity $\boldsymbol{t}_{inf}$ is selected as the predicted answer.

4. Experiments

Datasets and Evaluation Metrics. We conduct extensive experiments on two datasets: Outside Knowledge VQA (OK-VQA) [24] and Knowledge-Routed VQA (KRVQA) [7]. OK-VQA contains more than 14,000 questions that cover a variety of 10 knowledge categories. It is diverse and challenging since all questions are human-annotated without fixed question templates or knowledge bases, which require exploring a wide range of open-ended knowledge resource. We evaluate the performance by the standard VQA evaluation metric [3]. KRVQA [7] is to date the largest knowledge-based VQA dataset. It evaluates the multi-step reasoning ability of the models based on external knowledge. We use top-1 accuracy as in [7] for fair comparison.

Implementation Details. For all experiments, we train our model with PyTorch [27]. The softmax temperature τ in Eq. 3 is set to 1.0. We use all the annotated answers in the training set to construct knowledge triplets. For the triplet ranking loss, we treat all samples in a batch with different answers from the positive samples as negative samples. The margin is set to 1.0. Our model is trained by AdamW [19] optimizer with 200 epochs, where the batch size is 256 and the learning rate is set to 1×10^{-5} and 1×10^{-4} in the pre-training and fine-tuning stage, respectively.

4.1. Comparison with State-of-the-Art Methods

Comparison on OK-VQA: Table 1 shows the comparison results with state-of-the-art models, including knowledge graph based approaches [9, 41, 45, 46], unstructured-knowledge based approaches [24], multi-source knowledge based hybrid approaches [16,23], implicit knowledge based pre-training approaches [20,34], and the multimodal knowledge based approach [44]. Meanwhile, we also compare with traditional VQA methods [5, 14].

Our model MuKEA consistently outperforms all the

Method	KB-not-related							KB-related					Overall
	one-step			two-step				one-step	two-step				
	0	1	2	3	4	5	6	2	3	4	5	6	
Q-type [7]	36.19	2.78	8.21	33.18	35.97	3.66	8.06	0.09	0.00	0.18	0.06	0.33	8.12
LSTM [7]	45.98	2.79	2.75	43.26	40.67	2.62	1.72	0.43	0.00	0.52	1.65	0.74	8.81
FiLM [29]	52.42	21.35	18.50	45.23	42.36	21.32	15.44	6.27	5.48	4.37	4.41	7.19	16.89
MFH [43]	43.74	28.28	27.49	38.71	36.48	20.77	21.01	12.97	5.10	6.05	5.02	14.38	19.55
UpDn [2]	56.42	29.89	28.63	49.69	43.87	24.71	21.28	11.07	8.16	7.09	5.37	13.97	21.85
MCAN [42]	49.60	27.67	25.76	39.69	37.92	21.22	18.63	12.28	9.35	9.22	5.23	13.34	20.52
+ knowledge retrieval [7]	51.32	27.14	25.69	41.23	38.86	23.25	21.15	13.59	**9.84**	9.24	5.51	13.89	21.30
MuKEA	**59.12**	**44.88**	**37.36**	**52.47**	**48.08**	**35.63**	**31.61**	**17.62**	6.14	**9.85**	**6.22**	**18.28**	**27.38**

Table 2. State-of-the-art comparison on KRVQA dataset. The numbers in the third row mean different types of questions.

existing approaches and is superior to the state-of-the-art model [45] remarkably by 3.35%. Compared with most models following the 'knowledge retrieve and read' pipeline and referring to fixed knowledge bases, our end-to-end model effectively avoids cascading error while benefits from human-focused diverse multimodal knowledge. Moreover, our model greatly outperforms the pre-trained models by 10% since our model captures the question-centric and information-abstracted multimodal knowledge instead of simple vision and language co-occurrence 'knowledge' in the pre-training framework. Though KM^4 leverages multimodal knowledge by correlating images with entities in the existing knowledge graph, it still lacks of knowledge with high-order complex relationships and is inferior to MuKEA by 11.27%.

Comparison on KRVQA: In Table 2, we compare MuKEA with traditional VQA models [2, 29, 42, 43] and the knowledge-based model [7]. 'KB-not-related' questions only require basic visual knowledge while 'KB-related' questions need factual knowledge from knowledge bases.

Our model consistently outperforms existing models and achieves a remarkable boost of 6.08% on the overall metric over the best model [2]. It's worth to note that MuKEA obtains 7.81% improvement on average over [2] on the 'KB-not-related' questions, which indicates that even the vision-only questions require multimodal commonsense to bridge the low-level visual content and high-level semantics. MuKEA is inferior to some models on two-step reasoning *type 3* questions since that the answers of these questions are mostly relations while the accumulated and predicted tail entities of MuKEA are fact entities in most cases.

4.2. Ablation Study

In Table 3, we evaluate the contribution of knowledge learning losses, knowledge extraction schema, and knowledge accumulation strategy in MuKEA on the OK-VQA dataset. (1) In models '2-5', we evaluate the **effect of each loss function** on the performance. The accuracy of removing $\mathcal{L}_{\text{Tri}}$ and $\mathcal{L}_{\text{Sem}}$ respectively decreases by 1.24% and 0.53% while removing $\mathcal{L}_{\text{TransE}}$ results in a significant decrease in model '5'. Because $\mathcal{L}_{\text{TransE}}$ preserves the embedding structure of the whole triplets in our multimodal knowledge base, which has greater impact than $\mathcal{L}_{\text{Tri}}$ and $\mathcal{L}_{\text{Sem}}$. Model '4' results in a further decrease compared with '2' and '3', which indicates the complementary benefits of $\mathcal{L}_{\text{Tri}}$ and $\mathcal{L}_{\text{Sem}}$. (2) In models '6-8', we assess the **influence of triplet extraction methods**. For head entity extraction, we replace Gumbel-Softmax with soft attention in '6' and the performance drops by 1.92%. It's because that the head entity derived from LXMERT already contains object-centric contextual semantics for complex questions while directly fusing object features together introduces unexpected noise. Similarly, we apply self-attention over all the output tokens of LXMERT to represent the relation in '7' and the accuracy decreases by 1.80% compared with using [CLS] token, which benefits from the the pre-training classification task to contain highly-correlated multimodal

Method	**Accuracy**
1. MuKEA (full model)	**42.59**
Ablation of Loss Function	
2. w/o $\mathcal{L}_{\text{Tri}}$	41.35
3. w/o $\mathcal{L}_{\text{Sem}}$	42.06
4. w/o $\mathcal{L}_{\text{Tri}}$ & $\mathcal{L}_{\text{Sem}}$	40.84
5. w/o $\mathcal{L}_{\text{TransE}}$	24.50
Ablation of Triplet Representation	
6. head entity w/ soft-attention	40.67
7. relation w/ self-attention	40.79
8. tail entity w/ GloVe	41.42
Ablation of Triplet Structure	
9. w/o h	39.83
10. w/o r	39.40
Ablation of Knowledge Source	
11. w/o VQA 2.0 knowledge	36.35
12. w/o OK-VQA knowledge	27.20
Ablation of Pre-training Knowledge	
13. w/o LXMERT pre-training	33.52

Table 3. Ablation of key components in MuKEA on OK-VQA.

Method	Failure subset		
	MUTAN + AN*	Mucko*	KRISP*
MuKEA	40.09	40.06	40.46

(a)

Method	Failure subset
	MuKEA
MUTAN + AN*	26.45
Mucko*	27.68
KRISP*	27.68

(b)

Table 4. MuKEA accuracy on the failure subset of KB-based models (a) and vice versa (b). * indicates the model is re-implemented.

information. Furthermore, we utilize GloVe [28] to represent the tail entity in '8', resulting in a 1.17% accuracy drop because of the heterogeneous gap between fixed word embeddings and multimodal triplet representations. (3) In models '9-10', we prove the **importance of triplet structure**. We remove the head entity and the tail entity respectively. The performance drops 2.76% and 3.19% accordingly, which proves the effectiveness of our triplet-based knowledge organization structure. (4) In models '11-12', we prove the **importance of pre-training and fine-tuning strategy for knowledge accumulation**. It's obvious that without either of the two processes, the performance decreases remarkably. Though the basic knowledge in VQA 2.0 is less influential than the domain-specific knowledge in OK-VQA, the two working together achieves the best performance. (5) In model '13', we further test the **influence of prior knowledge accumulated in the pre-trained LXMERT**. The accuracy drops 9.07% without pre-training since both the head entity and the relation representations rely on the contextual information from the pre-trained knowledge.

4.3. Knowledge Complementary Analysis

To prove the complementary benefits of our multimodal knowledge with existing knowledge bases, we conduct two experiments on the OK-VQA dataset: (1) performance of MuKEA and existing models on mutual failure cases, and (2) performance of ensemble models of MuKEA and existing models. Here we test on three typical KB-based models: MUTAN + AN [24] on unstructured Wikipedia, Mucko [46] on structured ConceptNet, and KRISP [23] on multiple knowledge bases. We re-implemented these models for fair comparison on the same subset.

Table 4 shows the performance of MuKEA on the failure OK-VQA test subset of the above three models and vice versa. MuKEA consistently achieves over 40% accuracy on all the failure cases of the KB-based models (Table 4(a)). Meanwhile, KB-based models obtain over 26% accuracy on questions difficult for MuKEA (Table 4(b)). It proves that multimodal knowledge and existing KB knowledge respectively deals with different types of open-ended questions.

Method	Accuracy
MuKEA	42.59
MUTAN + AN*	25.43
MuKEA + (MUTAN + AN*)	35.39
MuKEA + (MUTAN + AN*) oracle	43.64
Mucko*	27.17
MuKEA + Mucko*	35.97
MuKEA + Mucko* oracle	44.84
KRISP*	32.02
MuKEA + KRISP*	37.75
MuKEA + KRISP* oracle	47.15

Table 5. Performance of model ensemble on OK-VQA.

Method	Accuracy	mAccuracy
KRISP*	32.31	26.91
MuKEA	42.59	35.42

Table 6. Long-tail analysis on OK-VQA dataset.

We further assemble MuKEA with three models respectively: if the difference of the top 2 minimum distances predicted by Eq. 10 is larger than a threshold m ($m = 0.07$), we select the predicted answer of MuKEA, otherwise, selecting the other. In Table 5, the baseline models are respectively improved by 9.96%, 8.80% and 5.73% after model ensemble. We also present the oracle setting that takes the accurate prediction from either of the models as the answer. The oracle performance obtains significant improvement, which further proves the complementary benefits of multimodal knowledge and existing knowledge bases.

4.4. Long-Tail Analysis

To prove the model's generalization ability on the rare answers while not overfitting on the 'head' ones, we propose a new unbiased metric mean Accuracy (**mAccuracy**) to fairly evaluate the performance on the long-tail distributed answers. Inspired by the unbiased metric in scene graph generation [8,35], mAccuracy calculates the accuracy for each unique answer separately and average for all the answers. We compare MuKEA with KRISP, which demonstrates its great generalization ability by referring to multiple knowledge sources. In Table 6, our model greatly outperforms KRISP by 8.51% on mAccuracy, which proves the strong generalization ability of the multimodal knowledge on the long-tail knowledge without sacrificing the accuracy of frequent referred knowledge.

4.5. Qualitative Analysis

From case study in Figure 3, we conclude that our model is interpretable by visualizing the predicted multimodal knowledge triplets: (1) **MuKEA captures instantiated knowledge beneficial for object understanding.** The

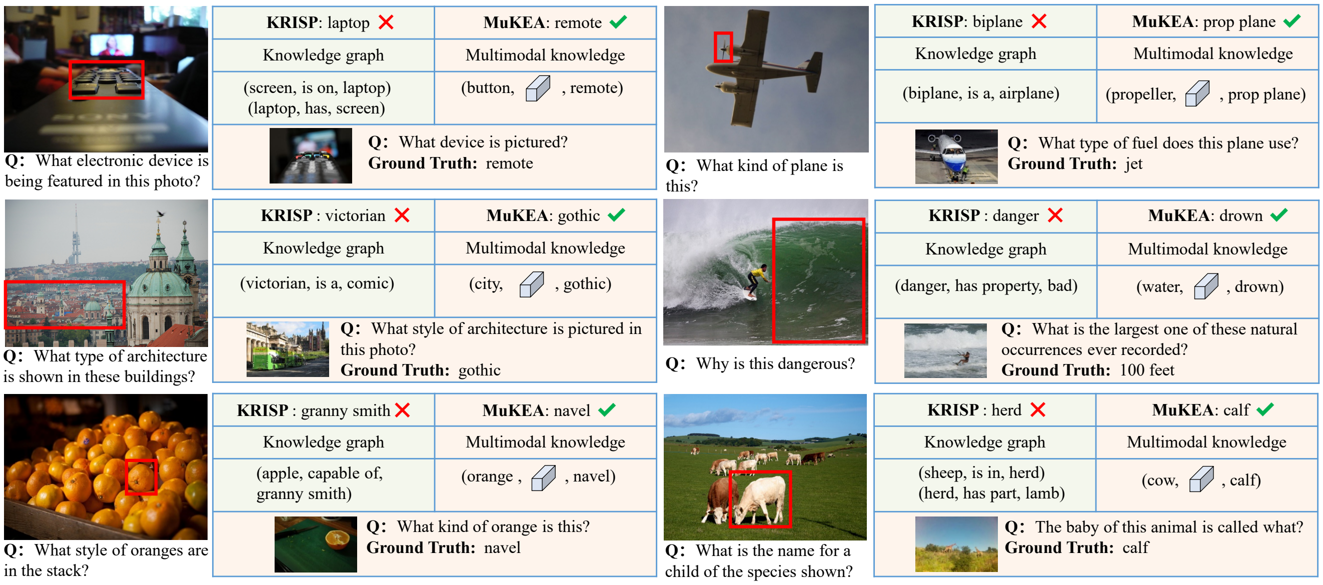
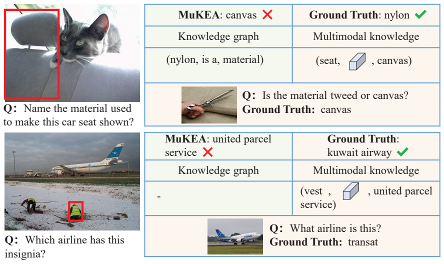


Figure 3. Visualization of the predicted answers and supporting knowledge of KRISP (green) and MuKEA (pink). For MuKEA, the red box in the image shows the head entity (α_i in Eq. 3). The bottom VQA training sample, which has the nearest relation embedding with the test sample, shows the scenario that accumulates relational knowledge supporting the current inference. The answer is shown in the tail.

examples in the first row indicate that MuKEA captures the complex knowledge between the object appearance and the object-centric facts. The supporting knowledge is in the form of an entire triplet (left example) or just the inexpressible relation (right example). (2) **MuKEA contains multi-object involved complex knowledge beneficial for scene understanding.** In the second row, MuKEA is capable to correlate the visual content of groups of buildings with the city style 'gothic'. (3) **MuKEA avoids the cascading error by directly reasoning on knowledge embeddings.** Existing models generally first detect object labels to retrieve relevant knowledge, which introduces unexpected noise with false labels. MuKEA has the advantage of adopting the semantic-rich embeddings to represent the knowledge and reason about the answer in an end-to-end mode.

4.6. Limitation Analysis

MuKEA fails mostly in the following cases (Figure 4): (1) MuKEA lacks adequate multimodal knowledge, such as the knowledge to distinguish *nylon* and *canvas*, due to the limited VQA scenarios in the training stage. (2) MuKEA fails in extracting some triplets. Since the head entities and the relations are extracted in an unsupervised mode, vision-similar content causes attention deviation, such as the vest is incorrectly attended as the *insignia*. The above problems need further research in accumulating more comprehensive knowledge and evaluating the triplet extraction quality. We also test the MuKEA on VQA 2.0 with inferior results to some works since that questions in VQA 2.0 mainly rely on visual appearance clues instead of external knowledge.



MuKEA: canvas ✗
Ground Truth: nylon ✓
Knowledge graph
Multimodal knowledge
(nylon, is a, material)
(seat, , canvas)
Q: Is the material tweed or canvas?
Ground Truth: canvas
Q: Name the material used to make this car seat shown?
MuKEA: united parcel service ✗
Ground Truth: kuwait airway ✓
Knowledge graph
Multimodal knowledge
-
(vest , , united parcel service)
Q: What airline is this?
Ground Truth: transat
Q: Which airline has this insignia?

Figure 4. Representative failure cases of MuKEA on OK-VQA.

5. Conclusion

In this paper, we propose a novel framework for knowledge-based visual question answering, which focuses on multimodal knowledge extraction and accumulation instead of using external knowledge bases. We propose a novel schema to represent multimodal knowledge by an explicit triplet and three loss functions to learn the representations from coarse to fine. We adopt a pre-training and fine-tuning strategy to accumulate multimodal knowledge progressively. Our model outperforms state-of-the-art on KB-VQA datasets and advances recent research from the multimodal knowledge view. We prove the complementary to existing knowledge graph. How to effectively combine MuKEA with knowledge bases will be the future work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62006222).

References

[1] Wikipedia: The free encyclopedia. https://www.wikipedia.org/. 1, 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 6

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 5

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. 1

[5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2612–2620, 2017. 5

[6] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013. 4

[7] Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2, 5, 6

[8] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 7

[9] François Garderes, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 489–498, 2020. 1, 2, 5

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 5

[11] Yudong Han, Yangyang Guo, Jianhua Yin, Meng Liu, Yupeng Hu, and Liqiang Nie. Focal and composed vision-semantic modeling for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4528–4536, 2021. 1

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference On Learning Representations*, 2017. 4

[13] Amar Viswanathan Kannan, Dmitriy Fradkin, Ioannis Akrotirianakis, Tugba Kulahcioglu, Arquimedes Canedo, Aditi Roy, Shih-Yuan Yu, Malawade Arnav, and Mohammad Abdullah Al Faruque. Multimodal knowledge graph for deep learning papers and code. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3417–3420, 2020. 2, 3

[14] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Proceedings of the 32rd International Conference on Neural Information Processing Systems*, 2018. 5

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[16] Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235, 2020. 2, 5

[17] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, 2020. 2, 3

[18] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 1, 2

[19] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *International Conference On Learning Representations*, 2018. 5

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23, 2019. 2, 5

[21] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019. 4

[22] Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107):6, 2004. 3

[23] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 1, 2, 5, 7

[24] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019. 2, 5, 7

[25] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution

nets for factual visual question answering. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018. 2

[26] Fudong Nian, Bing-Kun Bao, Teng Li, and Changsheng Xu. Multi-modal knowledge representation learning via webly-supervised relationships mining. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 411–419, 2017. 3

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019. 5

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 7

[29] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 6

[30] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3208–3218, 2018. 2, 3

[31] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International Semantic Web Conference*, pages 177–185. Springer, 2016. 2, 3

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015. 3

[33] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1405–1414, 2020. 2

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5100–5111, 2019. 2, 3, 5

[35] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 7

[36] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427, 2017. 2

[37] Peng Wang, Qi Wu, Chunhua Shen, Anthony R Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017. 1, 2

[38] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 3

[39] Jiawang Xie, Zhenhao Dong, Qinghua Wen, Hongyin Zhu, Hailong Jin, Lei Hou, and Juanzi Li. Construction of multi-modal chinese tourism knowledge graph. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 16–29. Springer, 2021. 2

[40] Jing Yu, Weifeng Zhang, Yuhang Lu, Zengchang Qin, Yue Hu, Jianlong Tan, and Qi Wu. Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(12):3196–3209, 2020. 1

[41] Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108:107563, 2020. 5

[42] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 6

[43] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018. 6

[44] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Km4: Visual reasoning via knowledge embedding memory model with mutual modulation. *Information Fusion*, 67:14–28, 2021. 2, 5

[45] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Knowledge is power: Hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2360–2368, 2021. 5, 6

[46] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1097–1103, 2020. 1, 2, 3, 5, 7