Boosting Visual Question Answering with Context-aware Knowledge Aggregation

Guohao Li
ligh16@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Xin Wang*
xin_wang@tsinghua.edu.cn
Tsinghua University
Beijing, China

Wenwu Zhu*
wwzhu@tsinghua.edu.cn
Tsinghua University
Beijing, China

ABSTRACT

Given an image and a natural language question, Visual Question Answering (VQA) aims at answering the textual question correctly. Most VQA approaches in literature targets at finding answers to the questions solely based on analyzing the given images and questions alone. Other works that try to incorporate external knowledge into VQA adopt a query-based search on knowledge graphs to obtain the answer. However, these works suffer from the following problem: the model training process heavily relies on the ground-truth knowledge facts which serve as supervised information — missing these ground-truth knowledge facts during training will lead to failures in producing the correct answers. To solve the challenging issue, we propose a Knowledge Graph Augmented (KG-Aug) model which conducts context-aware knowledge aggregation on external knowledge graphs, requiring no ground-truth knowledge facts for extra supervision. The proposed KG-Aug model is capable of retrieving context-aware knowledge subgraphs given visual images and textual questions, and learning to aggregate the useful image- and question-dependent knowledge which is then utilized to boost the accuracy in answering visual questions. We carry out extensive experiments to validate the effectiveness of our proposed KG-Aug models against several baseline approaches on various datasets.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; *Knowledge representation and reasoning*; • **Information systems** → *Question answering*.

KEYWORDS

visual question answering, knowledge graph

ACM Reference Format:
Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413943

*Corresponding Authors.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413943

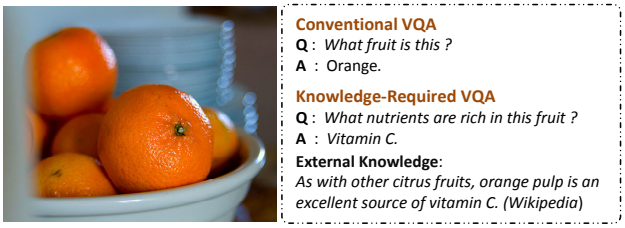

Figure 1: Examples showing the differences between conventional VQA and external knowledge-required VQA.

1 INTRODUCTION

With the rapid development of machine learning and its applications in computer vision, there is an increasing trend in the research community towards a better understanding of visual contents in order to achieve more general machine intelligence. As a well-documented cross-modal task, Visual Question Answering (VQA) provides a great opportunity for examining the comprehensive understanding of visual-language information, which has attracted a lot of attention from both academia and industry.

On the one hand, substantial efforts have been devoted to handling conventional VQA tasks that rely on low-level visual perceptions over the visual contents of given images. These tasks include verifying the existence of visual objects, detecting their locations in the images and recognizing their attributes, etc. Existing approaches on conventional VQA directly combine features of visual images and textual questions together using either attention mechanism [1, 15] or multimodal fusion techniques [3, 6, 12], requiring information only from the questions and images.

On the other hand, it is necessary even for a human to utilize external knowledge that cannot be directly inferred from the given visual images and textual questions to answer the questions accurately. As such, the abilities of capturing information from visual images and textual questions as well as learning to retrieve and utilize knowledge from external sources are indeed crucial for a VQA algorithm. Existing methods on external knowledge-required VQA either perform explicit queries on large-scale knowledge graphs (KG) [27] or retrieve the ground-truth knowledge facts (i.e., *subject-relation-object* triplets) from a close-domain knowledge base [17, 18, 26] to obtain the necessary knowledge, assuming that the querying results returned from the knowledge graph must contain the answers to the questions. Figure 1 gives an example of presenting the differences between conventional VQA and external knowledge-required VQA.

However, both of these two categories of approaches suffer from difficulties when bridging the given visual images and textual questions with the necessary external knowledge:

- The conventional VQA approaches lack the mechanism for incorporating external knowledge.
- The external knowledge-required VQA approaches require the ground-truth facts extracted from knowledge graphs as extra supervised information and will fail to obtain the correct answers when there are no available ground-truth facts (this could happen frequently) to supervise the model training or the referred questions do not need any external knowledge.

To tackle these difficulties, we propose a Knowledge Graph Augmented (KG-Aug) model which learns to conduct context-aware external knowledge aggregation on external knowledge graphs, requiring no ground-truth knowledge facts for extra supervision. Given the visual image and textual question, the proposed KG-Aug module first extracts KG entities/concepts that appear in the visual image or textual question, and construct a context-relevant knowledge subgraph with these extracted entities as anchor points. The anchor entities in subgraphs will then be encoded into context-aware vector forms, aggregating knowledge from their neighbor entities. To smoothly incorporate the aggregated knowledge into VQA approaches, we augment the feature representations of visual objects and textual questions with the vectorized knowledge representation through a context-aware deep fusion mechanism. Furthermore, we would like to point out that our proposed KG-Aug model can be combined with a wide range of VQA approaches which predict the answers by utilizing the feature representations of visual objects and textual words to boost their performances. To summarize, our work makes the following contributions:

- We boost VQA by proposing a Knowledge Graph Augmented (KG-Aug) model which overcomes the weaknesses of existing VQA approaches through performing context-aware knowledge aggregation on external knowledge graphs.
- We smoothly incorporate the aggregated external knowledge from large-scale knowledge graphs into VQA through a context-aware deep fusion mechanism that requires no ground-truth facts as supervised information.
- Our proposed model is able to boost the accuracy of various VQA approaches that utilize the feature representations of visual objects and textual words to predict the answer.
- We conduct extensive experiments to provide promising results demonstrating the effectiveness of the proposed KG-Aug model against several state-of-the-art approaches on various datasets.

2 RELATED WORK

Answering conventional visual questions. We refer to the conventional visual questions as the questions that are answerable merely from visual contents. For the conventional visual questions, the widely adopted solution is *jointly embedding* the visual and question features into a common space using advanced attention mechanism [1, 15] or multimodal fusion techniques [3, 6, 12], then feeding them into a classifier over candidate answers. In recent years, this type of method has dominated the VQA open challenge and witnessed an impressive accuracy boosts (over 75% accuracy in VQA2.0 benchmark dataset [9] today).

Encoding unstructured text as external knowledge. Several previous works [7, 16, 28] make use of unstructured text corpus (e.g., Wikipedia text, natural language sentences) as the source of external knowledge. In these methods, external text information is usually encoded (using techniques such as Doc2Vec, Recurrent Neural Networks, BERT etc.) into feature vectors to serve as external knowledge. The difficulty lies in the fact that unstructured text usually contains much noisy information and it is usually beyond these models' capabilities to jointly learn knowledge from the external text corpus and leverage the knowledge to answer questions.

Explicit reasoning on knowledge graphs. Compared with unstructured text, large-scale knowledge graphs provide clear and structured information, which covers knowledge ranging from commonsense to topic-specific and even expert knowledge in the form of graphs. Substantial efforts have been devoted to building large-scale knowledge graphs [2, 22, 25] in recent years. A major challenge here is the underlying heterogeneity between symbolic knowledge graphs and continuous visual signals. Several early works [27] choose to unify them together in the symbolic space. These methods describe visual information with symbols (e.g., visual concepts) and link them to relevant parts of knowledge graphs, in order to conduct explicit reasoning along the graph edges. However, these methods largely depend on the pre-defined templates for reasoning and fail to preserve rich visual information.

Fact retrieval from knowledge bases. Different from previous methods, several works [17, 18, 26] formulate this problem as a retrieval task that localizes the most relevant *facts* (i.e., *subject-relation-object* triplets) from the given knowledge base, where a learnable score function is usually applied to model the relevance between the question-image pairs and candidate facts. The final answer is selected from the entities that appear in the retrieved facts. These retrieval-based methods exhibit superior robustness and versatility compared with the explicit reasoning methods. However, these methods largely depend on ground-truth knowledge facts which serve as extra supervised information and may fail when there are no available ground-truth knowledge facts or the referred questions do not need any external knowledge.

Recently, there are also related works that expoit retrieved knowledge facts throught memory networks [14, 20] or graph neural networks [21]. In contrast, our work performs knowledge aggregation based on KG entities and can be smoothly intergrated with the conventional VQA models to boost their performance.

3 VQA WITH CONTEXT-AWARE KNOWLEDGE AGGREGATION

In this section, we elaborate on the details of our proposed Knowledge Graph Augmented (KG-Aug) model that learns to conduct context-aware external knowledge aggregation and smoothly incorporate the aggregated knowledge into VQA models through a context-aware deep fusion mechanism to boost their performance. Figure 2 gives an overview of our model based on a visualized example.

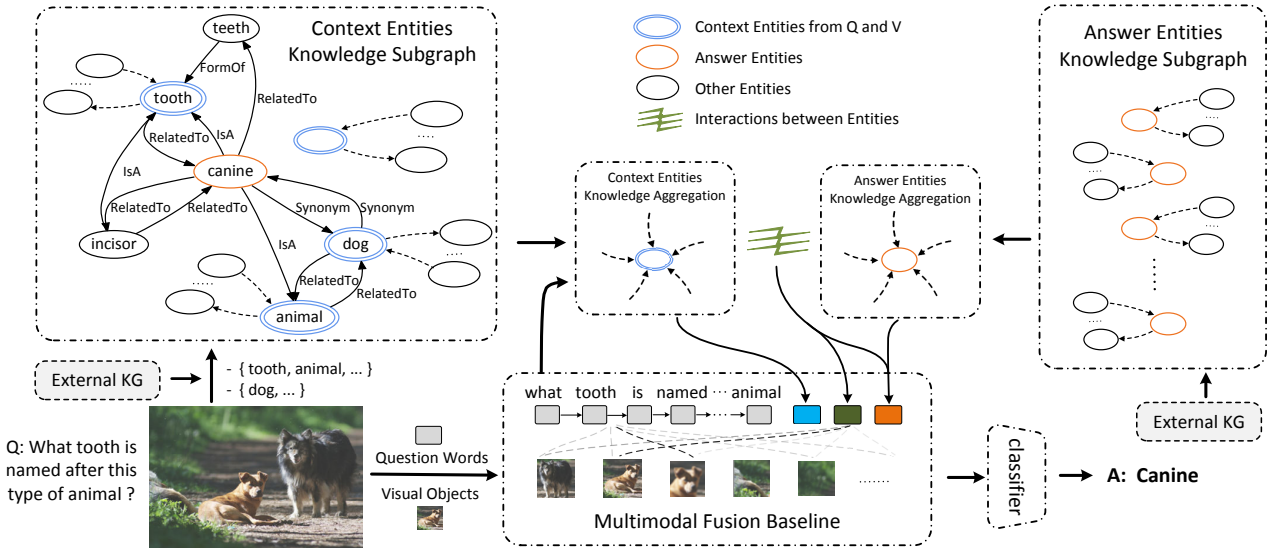

Figure 2: Overview of our proposed KG-Aug model. Given a pair of image and question, our model first retrieves context-aware knowledge subgraph from external large-scale KGs, then conducts a context-aware knowledge aggregation on the subgraphs to embed the knowledge into several anchor entities. By jointly modeling the context entities and candidate answer entities, we encapsulate external knowledge into three supplementary features, which are then smoothly incorporated into a fusion-based VQA model to boost its performance.

Technically, augmenting conventional VQA models with external large-scale knowledge graphs is non-trivial. The difficulties arise from the three aspects:

(1) There is a discrepancy between the overwhelming amount of available external knowledge and the restricted question-image context information. We close this discrepancy by retrieving a context-aware knowledge subgraph in Sec. 3.1.
(2) According to visual and textual context, it is challenging to learn how to aggregate useful information from knowledge graphs. We leverage the graph convolution techniques to aggregate knowledge into several anchor entities in Sec. 3.2.
(3) Another challenge lies in that, how to smoothly fuse aggregated knowledge into VQA systems, so that the system can perform well on various questions. We describe the fusion schemes in Sec. 3.3.

3.1 Knowledge Subgraph Retrieval

It is a preliminary preparation to retrieve an appropriate amount of knowledge from large-scale knowledge graphs (KGs) and generate a *context-aware knowledge subgraph*. In this work, we exploit external knowledge from two large-scale knowledge graphs: 1) ConceptNet [22] that contains commonsense relationships between daily words and phrases that people use; and 2) Wikidata [25] that provides extensive factual knowledge about our world.

Specific to a question-image context, an ideal retrieval procedure would retrieve potentially useful information in this context and ignore the irrelevant ones, which is important to reduce the computational costs and prevent data noises from misleading the model. To this end, we adopt a two-step procedure to retrieve a knowledge subgraph: (1) We associate key entities appearing in the context (image and question, etc.) to the large-scale KGs and denote them as *anchor entities*; (2) We build a *context-aware knowledge subgraph* by exploiting the anchor entities and their neighborhoods in the large-scale KGs.

In detail, we collect two types of anchor entities from the context, including textual and visual entities. The textual entities are KG entities associated with key phrases of the question, where we utilize the Stanford NLP Dependency Parser [4] and Stanford Named Entity Recognizer [5] to analyze the question and extract all noun phrases, verb-object phrases and named entities as the key phrases. The visual entities are KG entities associated with visual objects in the image, where we extract the names of prominent visual objects with a Faster-RCNN object detector [8] pre-trained on Visual Genome dataset [13]. These anchor entities establish connections between the visual-question context and external large-scale KGs (ConceptNet and Wikidata).

Afterward, we reduce the large-scale KGs into a context-aware subgraph to preserve the most useful candidate knowledge and discard irrelevant ones. Taking the anchor entities as starting points in external KGs, we perform an expansion from the anchor entities to their first-order neighbors and preserve the neighborhood entities and edges which finally constitute the subgraph.

In addition to the *local* knowledge subgraph for each question-image sample, we also build a *global* knowledge subgraph for all candidate answers, whose anchor entities are all possible answers (maybe thousands). The motivation of building a global answer knowledge subgraph is to model answer labels as answer entities, in a similar manner as we model visual entities and textual entities. It enables us to bridge questions, images and candidate answers together in a common knowledge-based semantic space.

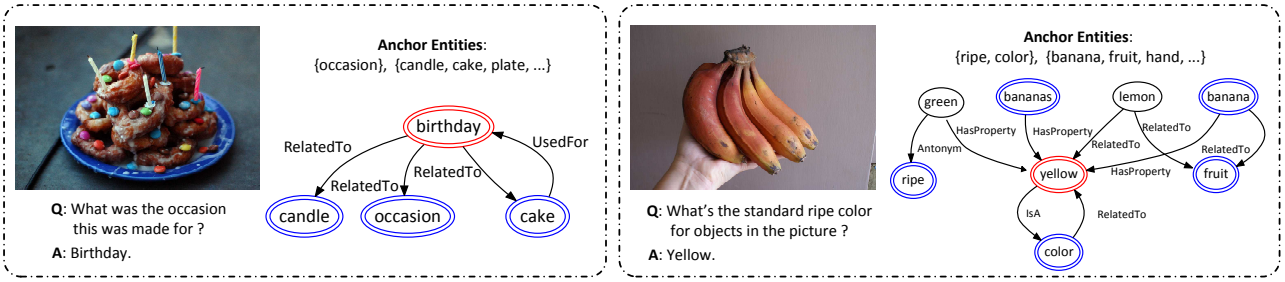

Figure 3: Two real examples in our experimental datasets. We visualize the expected knowledge subgraphs that are potentially useful for answering these visual questions.

In the end, the subgraphs act as the context-aware external knowledge to be exploited and incorporated in VQA.

3.2 Aggregating Knowledge into Anchor Entities

Given the knowledge subgraph we have built, it remains challenging to represent the graph in accordance with the images and questions so that they are computable in a unified and scalable system. Generally, the most effective conventional VQA approaches [1, 10, 24] are built upon hidden representations of images and questions, which motivates us to embed knowledge graph into implicit representations. Inspired by recent works on graph neural networks [30] and knowledge-based question answering [29], we utilize a particular type of graph convolution network to aggregate graph knowledge into anchor entities according to a specific contextual query.

Contextual Query Generation. We consider the input question as a contextual query that indicates what external knowledge might be useful. For a question q with M words, we first embed each word into GloVe vectors [19], then feed them sequentially into a multi-layer LSTM to get a sequence of hidden states $\{\boldsymbol{h}_m^q\}_{m=1,\cdots,M}$. We take the final LSTM hidden states $\boldsymbol{h}^q$ as the contextual query vector.

Entity Neighborhood Scoring. Each entity e is linked to neighboring entities e_i through relations r_i which constitute a neighborhood $\mathcal{N}_e$ for entity e. The triplet (e, r_i, e_i) represents a piece of knowledge, and can be of different importance according to different contextual queries.

In order to model the importance of neighboring relations and entities (r_i, e_i) for an anchor entity e under a contextual query vector $\boldsymbol{h}^q$, we take two aspects into consideration: (1) similarity between the query vector and relation and (2) whether the neighboring entity is also an anchor entity. Intuitively, an edge connecting two anchor entities should be more relevant than other edges for answering questions.

The importance score for neighboring relations and entities (r_i, e_i) is computed as follows:

$$s_{(r_i,e_i)} \propto \exp\left(\mathrm{I}\left[e_i \in \mathcal{E}_0\right] + \vec{\boldsymbol{r}}_i \cdot \boldsymbol{h}^q\right), \tag{1}$$

where I is a binary indicator, $\mathcal{E}_0$ is the set of anchor entities and $\vec{\boldsymbol{r}}_i$ is a trainable embedding vector of r_i.

Aggregating Knowledge from Neighbors. At this stage, we aggregate knowledge from neighbors to update the embedding of anchor entities. Starting with a collection of initial entity embedding $\{\vec{\boldsymbol{e}}^{(0)}\}$, this updating operation can be done iteratively for up to T rounds.

At round t, we model the neighboring knowledge by feeding the neighboring entity embedding $\vec{\boldsymbol{e}}_i^{(t-1)}$ and relation embedding $\vec{\boldsymbol{r}}_i$ through a non-linear fully-connected layer. Then, the neighboring knowledge is weighted by their importance score and aggregated into a neighboring knowledge vector $\vec{\boldsymbol{e}}_{\mathcal{N}}^{(t)}$ as follows:

$$\vec{\boldsymbol{e}}_{\mathcal{N}}^{(t)} = \sum_{(e_i,r_i)\in\mathcal{N}_e} s_{(r_i,e_i)} \tanh\left(\mathrm{W}_1[\vec{\boldsymbol{r}}_i; \vec{\boldsymbol{e}}_i^{(t-1)}]\right). \tag{2}$$

After that, the aggregated knowledge is propagated to anchor entities through a gating mechanism, and the embedding of anchor entities can be updated as follows:

$$\gamma_e^{(t)} = \sigma\left(\mathrm{W}_2[\vec{\boldsymbol{e}}^{(t-1)}; \vec{\boldsymbol{e}}_{\mathcal{N}}^{(t)}]\right), \tag{3}$$

$$\vec{\boldsymbol{e}}^{(t)} = (1-\gamma_e^{(t)})\vec{\boldsymbol{e}}^{(t-1)} + \gamma_e^{(t)}\vec{\boldsymbol{e}}_{\mathcal{N}}^{(t)}, \tag{4}$$

where the $\gamma_e^{(t)}$ is a learned trade-off factor indicating how much neighboring knowledge is propagated into $\vec{\boldsymbol{e}}^{(t)}$.

After T rounds, the model yields a collection of embedding of anchor entities $\{\vec{\boldsymbol{e}}\} \equiv \{\vec{\boldsymbol{e}}^{(T)}\}$.

In summary, we exploit the knowledge subgraph by encoding the information into *embedding of anchor entities*, which makes it more feasible to incorporate the knowledge in the next stage.

3.3 Answer Visual Questions with Knowledge

With the knowledge graph embedding method described in Sec. 3.2 being applied to the two types of knowledge subgraphs extracted in Sec. 3.1, we obtain two collections of vectorized entities: a collection of *context entities* that encode external knowledge for image-question context, and a collection of *answer entities* that represent each corresponding answer as entities with semantic knowledge.

In this section, we turn to the challenge of how to seamlessly incorporate these anchor entity embeddings into conventional VQA approaches. Our model adopts the following two measures:

- Aggregating the entity embeddings into several **auxiliary question features** that encapsulate useful external knowledge, which further removes noisy information and reduces the scale of knowledge involved.
- Matching entities to visual regions to **augment visual features with semantic knowledge**.

By means of manipulating the question and visual features, we can seamlessly incorporate the embedded knowledge into conventional VQA pipelines.

Generating Auxiliary Question Features. We distill the embedded knowledge through three auxiliary features $\{\vec{\boldsymbol{e}}^{(\text{ctx})}, \vec{\boldsymbol{u}}, \vec{\boldsymbol{e}}^{(\text{ans})}\}$, which respectively correspond to: 1) the question-image context, 2) the context-answer compatibility and 3) the targeted answer that is the most compatible with the context.

We denote C as the set of context entities and $\mathcal{A}$ as the set of candidate answer entities. Given a context query vector $\boldsymbol{h}^q$ as we used in Sec. 3.2, we measure the similarity between $\boldsymbol{h}^q$ and each context entities $e_i \in C$, then combine them into a **contextual entity embedding** $\vec{\boldsymbol{e}}^{(\text{ctx})}$ as follows:

$$\vec{\boldsymbol{e}}^{(\text{ctx})} = \sum_{e_i \in C} \boldsymbol{\alpha}_i \vec{\boldsymbol{e}}_i, \quad \boldsymbol{\alpha}_i \propto \exp\left(\boldsymbol{h}^q \cdot \vec{\boldsymbol{e}}_i\right). \tag{5}$$

The formula $\boldsymbol{\alpha}_i \propto exp(*)$ means that we use a softmax operation to obtain the weights $\boldsymbol{\alpha}_i$, where the sum of $\boldsymbol{\alpha}_i$ equals to 1.

Then we take the answer entities into consideration. The merits of modeling answer labels as answer entities is that, it allows us to assess the compatibility between the context and the candidate answers in a common semantic space. For each candidate answer entity $e_j \in \mathcal{A}$, we compute its compatibility factor $\boldsymbol{\beta}_j$ with the context by computing the inner product between the embedding of itself $\vec{\boldsymbol{e}}_j$ and the previous contextual entity embedding $\vec{\boldsymbol{e}}^{(\text{ctx})}$:

$$\boldsymbol{\beta}_j \propto \exp\left(\vec{\boldsymbol{e}}^{(\text{ctx})} \cdot \vec{\boldsymbol{e}}_j\right). \tag{6}$$

Notably, higher compatibility usually indicates there are stronger connections (e.g., more reliable reasoning paths) between the context and the answer entity in their knowledge subgraph neighborhood. Thus, the above compatibility factors $\boldsymbol{\beta} \in \mathcal{R}^{|\mathcal{A}|}$ can be interpreted as the "closeness" between the context and each candidate answer in the knowledge graph space.

For ease of use, we further encode the context-answer compatibility factors $\boldsymbol{\beta} \in \mathcal{R}^{|\mathcal{A}|}$ into a **context-answer compatibility vector** $\vec{\boldsymbol{u}}$ through a non-linear fully-connected layer as follows:

$$\vec{\boldsymbol{u}} = \text{ReLU}\left(\boldsymbol{\beta} \cdot \text{W}_3\right), \tag{7}$$

where the W_3 is a trainable parameteric matrix.

Meanwhile, a **targeted answer entity embedding** $\vec{\boldsymbol{e}}^{(\text{ans})}$ is thereby calculated according to the compatibility factors as follows:

$$\vec{\boldsymbol{e}}^{(\text{ans})} = \sum_{e_j \in \mathcal{A}} \boldsymbol{\beta}_j \vec{\boldsymbol{e}}_j. \tag{8}$$

In the end, we obtain three features $\{\vec{\boldsymbol{e}}^{(\text{ctx})}, \vec{\boldsymbol{u}}, \vec{\boldsymbol{e}}^{(\text{ans})}\}$ that encapsulate useful external knowledge as auxiliary information for answering the questions. Altogether, these auxiliary features capture underlying reasoning paths in the graph and serve as essential components to make full use of the aggregated knowledge.

Augmenting Visual Features with Visual Entities. As for the visual side, we augment the visual features with their corresponding visual entities.

It should be noted that, we use object-based visual features the same as the *Bottom-up Top-down* model [1], which employs a Faster-RCNN [8] object detector as the feature extractor. Therefore, it allows us to link the visual features with semantic names and knowledge graph entities. In our work, we establish the correspondence by simultaneously extracting the visual features and predicting their class names.

Suppose an image has K visual features, where the k_{th} feature $\boldsymbol{v}_k$ corresponds to the context entity $e_k \in C$, we augment the visual features by concatenating $\boldsymbol{v}_k$ and $\vec{\boldsymbol{e}}_k$ together, producing the **augmented visual features** as:

$$\tilde{\boldsymbol{v}}_k = \left[\boldsymbol{v}_k; \vec{\boldsymbol{e}}_k\right]. \tag{9}$$

To some extent, this concatenation operation adds extra semantic information to the visual features, making them more expressive when interacting with questions features.

Incorporating Knowledge into VQA Pipelines. Here we elaborate on how to incorporate the obtained knowledge into VQA pipelines.

In general, conventional VQA approaches adopt a fusion operation to combine visual features and question features together, i.e., $\text{BaseFusion}(\{\boldsymbol{v}_k\}_{k=1,\cdots,K}; \{\boldsymbol{h}^q_m\}_{m=1,\cdots,M})$. In this work, we replace $\boldsymbol{v}$ with augmented visual features $\tilde{\boldsymbol{v}}$ and design two schemes to augment the three auxiliary features $\{\vec{\boldsymbol{e}}^{(\text{ctx})}, \vec{\boldsymbol{e}}^{(\text{ans})}, \vec{\boldsymbol{u}}\}$ to the question.

Scheme 1: Late Augmentation. The auxiliary features are regarded as extra question hidden features and concatenated with $\{\boldsymbol{h}^q_m\}$ as:

$$\vec{\boldsymbol{f}} = \text{BaseFusion}(\{\tilde{\boldsymbol{v}}_k\}; \{\vec{\boldsymbol{e}}^{(\text{ctx})}, \vec{\boldsymbol{e}}^{(\text{ans})}, \vec{\boldsymbol{u}}\} \cup \{\boldsymbol{h}^q_m\}). \tag{10}$$

Scheme 2: Early Augmentation. The auxiliary features are regarded as extra question words and fed into a LSTM to obtain $M+3$ question hidden states $\{\tilde{\boldsymbol{h}}^q_m\}_{m=1,\cdots,M+3}$ as follows:

$$\{\tilde{\boldsymbol{h}}^q_m\} = \text{LSTM}(\{\vec{\boldsymbol{e}}^{(\text{ctx})}, \vec{\boldsymbol{e}}^{(\text{ans})}, \vec{\boldsymbol{u}}\} \cup \{\boldsymbol{w}_m\}) \tag{11}$$

$$\vec{\boldsymbol{f}} = \text{BaseFusion}(\{\tilde{\boldsymbol{v}}_k\}; \{\tilde{\boldsymbol{h}}^q_m\}), \tag{12}$$

where the M is the question length, $\{\boldsymbol{w}_m\}$ is the sequence of question word embeddings.

In the end, the produced fusion vector $\vec{\boldsymbol{f}}$ is fed into a classifier (usually implemented as a multi-layer perceptron) to get the final answer, i.e., $\hat{\text{a}} = \text{Classifier}(\vec{\boldsymbol{f}})$.

We close this section by pointing out that, our KG-Aug model is a practical and generic framework to incorporate large-scale external knowledge graphs into VQA pipelines. It learns to aggregate context-aware external knowledge without extra supervision except for the answer and can be applied to a wide range of VQA baselines.

Table 1: Results on the OK-VQA dataset. We mark the results reported in previous papers with star symbols(*). The other results are calculated by averaging nine repeated experiments with casually selected random seeds to reduce the experiment variance. We show the results for the overall OK-VQA dataset and for each question category: Vehicles and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and Food (CF); Geography, History, Language and Culture (GHLC); People and Everyday Life (PEL); Plants and Animals (PA); Science and Technology (ST); Weather and Climate (WC); and Other.

Method	OKVQA	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
Q-Only* [16]	14.93	14.64	14.19	11.78	15.94	16.92	11.91	14.02	14.28	19.76	25.74	13.51
MLB-Att [12]	20.40	20.00	16.24	17.86	24.61	22.39	17.67	17.62	20.67	18.49	31.27	17.55
BAN* [11]	25.17	23.79	17.67	22.43	30.58	27.90	**25.96**	20.33	25.60	20.95	**40.16**	22.46
BAN (re-run) [11]	25.46	23.65	21.07	**22.55**	33.64	26.95	21.33	21.07	26.60	19.42	36.12	22.64
ArticleNet* [16]	5.28	4.48	0.93	5.09	5.11	5.69	6.24	3.13	6.95	5.00	9.92	5.33
BAN + ArticleNet* [16]	25.61	24.45	19.88	21.59	30.79	**29.12**	20.57	21.54	26.42	**27.14**	38.29	22.16
KG-Aug (ours)	16.11	15.54	12.11	13.94	19.60	16.93	14.61	13.37	16.55	11.48	27.92	14.98
MLB + KG-Aug (ours)	20.89	20.08	16.45	17.71	26.68	22.27	16.78	18.03	21.70	18.60	32.58	17.91
BAN + KG-Aug (ours)	**26.71**	**24.65**	**21.59**	22.42	**34.75**	28.67	23.97	**21.97**	**27.75**	23.28	38.85	**24.29**

4 EMPIRICAL EXPERIMENTS

We conduct experiments on three VQA datasets: (1) the OK-VQA (**O**utside **K**nowledge VQA) dataset [16], the largest human-written knowledge-based VQA dataset with more than 14K open-ended visual questions, which require both the commonsense and factual open knowledge to answer; (2) the Visual7W-telling dataset [31] that contains 328K multi-choice visual questions of various types *(What, Where, When, Who, Why and How)*, where lots of diverse questions involve external commonsense knowledge; and (3) the FVQA dataset [26] where each question-answer sample is accompanying a piece of ground-truth fact retrieved from a given knowledge base.

For the OK-VQA and Visual7W datasets, our Knowledge Graph Augmented (KG-AUG) model obtains state-of-the-art results. It generalizes well across different baseline approaches, demonstrating the effectiveness and universality of our model. For the FVQA dataset, we experiment without using the given knowledge base or the provided ground-truth facts, and the model can achieve consistent improvements over the baselines in this dataset.

We describe the baseline approaches and our model variations in Sec. 4.1 and analysis their performance on these datasets in Sec. 4.2. To further gain insights into our method, we carry out extensive ablation studies and provide visualized results in Sec. 4.3.

4.1 Baselines and Model Variations

In this section, we briefly describe the baseline approaches for comparison in this work, including conventional fusion-based models and knowledge-based ones. Besides, we introduce several model variations of ours, including a *standalone KG-Aug model* and several *X+KG-Aug models* that augment the baseline approach *X* with our *KG-Aug* model.

MLB-Att. A Multimodal Low-rank Bilinear model [12] with the attention mechanism. This model builds an attention distribution for the visual features based on the question using the low-rank bilinear fusion, which is a classic and widely-used feature fusion technique.

BAN. The Bilinear Attention Networks [11]. This model considers the bilinear interactions between each pair of input channels for visual and language features. It is one of the current state-of-the-art fusion-based VQA methods.

ArticleNet. A knowledge-based method that tries to incorporate external information from Wikipedia articles [16]. It uses a GRU to encode sentences in the articles and predicts whether each word in the article is the answer or not.

BAN + ArticleNet. A knowledge-based method that combines the BAN and the ArticleNet model [16]. The hidden states of ArticleNet are incorporated into the BAN model through a memory network [23].

KDMN. A knowledge-based method [14] that incorporates external knowledge facts with dynamic memory networks. It is designed for the multi-choice VQA tasks.

KG-Aug (ours). The standalone version of our KG-Aug model. We obtain the answer by feeding the context-answer compatibility factors $\boldsymbol{\beta} \in \mathcal{R}^{|\mathcal{A}|}$ (in Sec. 3.3) to a classifier, i.e., $\hat{a} = \text{Classifier}(\boldsymbol{\beta})$.

MLB + KG-Aug (ours) and **BAN + KG-Aug (ours).** The *X + KG-Aug* model takes a fusion-based VQA baseline and augments it with our proposed KG-Aug model. In our experiments, we take two representative models (the MLB-Att model and the BAN model) as the *X* baselines. Specifically, the *BAN+KG-Aug* model uses the *Late Augmentation Scheme* to incorporate knowledge and the *MLB+KG-Aug* model uses the *Early Augmentation Scheme*, which have been described in Sec. 3.3.

4.2 Performance Analysis

We conduct extensive experiments to assess the model's capability for answering knowledge-required visual questions (the OK-VQA dataset, Sec. 4.2.1 and the FVQA dataset, Sec. 4.2.3) and general visual questions with high diversity (the Visual7W dataset, Sec. 4.2.2). In this section, we report and analyze the experimental results on these datasets. The results demonstrate that our model exhibits superior performance and generalizes well under various experimental settings.

4.2.1 Results on the OK-VQA Dataset. For the open-ended OK-VQA dataset, we obtain the final answer â by a classification over all possible answers. We follow the original dataset paper [16] and use the standard VQA accuracy metric $\text{Acc}(\hat{a}) = \min\{\frac{\#\text{humans that said } \hat{a}}{3}, 1\}$ to evaluate the model performance.

As shown in Table 1, we provide the results of several fusion-based baseline models in the top rows, a knowledge-based *ArticleNet* baseline in the middle rows, and our models in the bottom rows. In order to allow meaningful comparisons, we report two *BAN* model results: one is the result reported in the previous paper [16] and the other one is our re-run result via repeated experiments with random seeds.

In comparison with the *BAN* baseline, our *BAN+KG-Aug* model improves the overall accuracy by 1.25%, achieving the best performance among all models. As for the *MLB-Att* baseline, our *MLB+KG-Aug* model gains 0.5% improvements over it. The consistent performance improvements indicate that our method can successfully adapt to different base models and is helpful for answering knowledge-required visual questions.

Compared with another knowledge-based *ArticleNet* method, our standalone model *KG-Aug* surpasses it with more than 200% relative performance boosts. Besides, we can measure the separate contributions of ours and the ArticleNet (i.e., *BAN+KG-Aug* and *BAN+ArticleNet*) to the *BAN* baseline model. Experiment results demonstrate that our method is better (1.25% v.s. 0.4%) at incorporating external knowledge.

Table 2: Results on Visual7W dataset.

Methods	Overall	What	Where	When	Who	Why	How
LSTM-Att* [31]	54.3	51.5	57.0	75.0	59.5	55.5	49.8
MCB + Att* [6]	62.2	60.3	70.4	79.5	69.2	58.2	51.1
KDMN* [14]	66.0	64.6	73.1	81.3	73.9	64.1	53.3
BAN [11]	71.1	71.3	76.7	82.6	78.1	64.4	59.3
KG-Aug (ours)	67.6	60.7	69.2	78.7	71.4	61.4	**84.9**
BAN + KG-Aug (ours)	**72.0**	**72.0**	**77.8**	**83.4**	**78.9**	**66.9**	59.6

4.2.2 Results on the Visual7W Dataset. In contrast with the previous open-ended dataset, the Visual7W dataset is a VQA dataset in the multi-choice setting with four candidate answers for each question. We formulate this multi-choice problem as a binary classification problem and predict a probability $P(c_i)$ for the i_th candidate answer, then make the final choice by selecting the one with highest probability $\hat{a} = \text{argmax}_{i\in\{1,2,3,4\}} P(c_i)$. The evaluation metric is the accuracy of making the correct choices. There are some minor modifications to our models for adapting to this dataset, such as concatenating the multi-choice answers with the question as input, etc.

In Table 2, we list several baselines for comparison. As expected, our *BAN+KG-Aug* model gains noticeable performance improvements (0.9%) compared to the *BAN* baseline, indicating that our model adapts well on this dataset. Our model obtains different boosts for different question categories, where the most significant improvement is for the *Why* questions (2.5%), due to the fact that *Why* questions usually require more commonsense knowledge to answer.

Table 3: Results on the FVQA dataset.

Methods	Accuracy
LSTM-Question+Image* [26]	22.97
Hie-Question+Image* [15]	33.70
BAN [11]	35.69
KG-Aug (ours)	31.96
BAN + KG-Aug (ours)	**38.58**

4.2.3 Results on the FVQA Dataset. The FVQA dataset is an open-ended VQA dataset similar to the OK-VQA. Besides, it also provides an accompanying knowledge base of facts and ground-truth fact annotation for each question-answer pair. We experiment without using the given knowledge base or the provided ground-truth facts in FVQA dataset, aiming to leverage open knowledge to boost the performance, only using the answer as training signals. Thus we did not compare with these fact retrieval based methods [17, 18]. We evaluate the model performance by using the standard accuracy metric and averaging among the five official train/val splits. As shown in Table 3, our *BAN+KG-Aug* model gains significant performance improvements (2.89%) compared to the *BAN* baseline, Besides, even the standalone *KG-Aug* model that only exploits the context-answer compatibility factors achieves competitive results. These experimental results indicate the effectiveness of our method for incorporating external knowledge.

4.3 Ablation Studies and Visualizations

To get further insights into our model, we conduct several ablation studies on the OK-VQA dataset to investigate the effects of the model components and the effectiveness of the knowledge sources (Table 4). In Fig. 4, we visualize examples to provide more insights.

Table 4: Ablation results on the OK-VQA dataset.

Ablation	Accuracy	Acc.Drops
- Full model *(BAN + KG-Aug)*	26.71	0.0
a) - w/o question augmentation		
- w/o $\{\vec{u}\}$	26.50	-0.21
- w/o $\{\vec{e}^{(\text{ans})}\}$	26.43	-0.28
- w/o $\{\vec{u}, \vec{e}^{(\text{ans})}\}$	25.52	-1.19
- w/o $\{\vec{e}^{(\text{ctx})}, \vec{u}, \vec{e}^{(\text{ans})}\}$	25.92	-0.79
- w/o visual augmentation	26.53	-0.18
- w/o knowledge aggregation	24.94	-1.77
b) - w/o knowledge sources		
- w/o ConceptNet	26.33	-0.38
- w/o Wikidata	26.49	-0.22
- with noisy knowledge	25.95	-0.76
- Base model *(BAN)*	25.46	-1.25

Ablation Studies # 1: Effects of Model Components. We measure the effects of each part in our KG-Aug model by evaluating the model accuracy when the part gets removed.

As shown in Table 4(a), we observe a significant accuracy decrease when removing the $\{\vec{u}, \vec{e}^{(\text{ans})}\}$ features, especially when they both get removed. We think the reason why the $\{\vec{u}, \vec{e}^{(\text{ans})}\}$

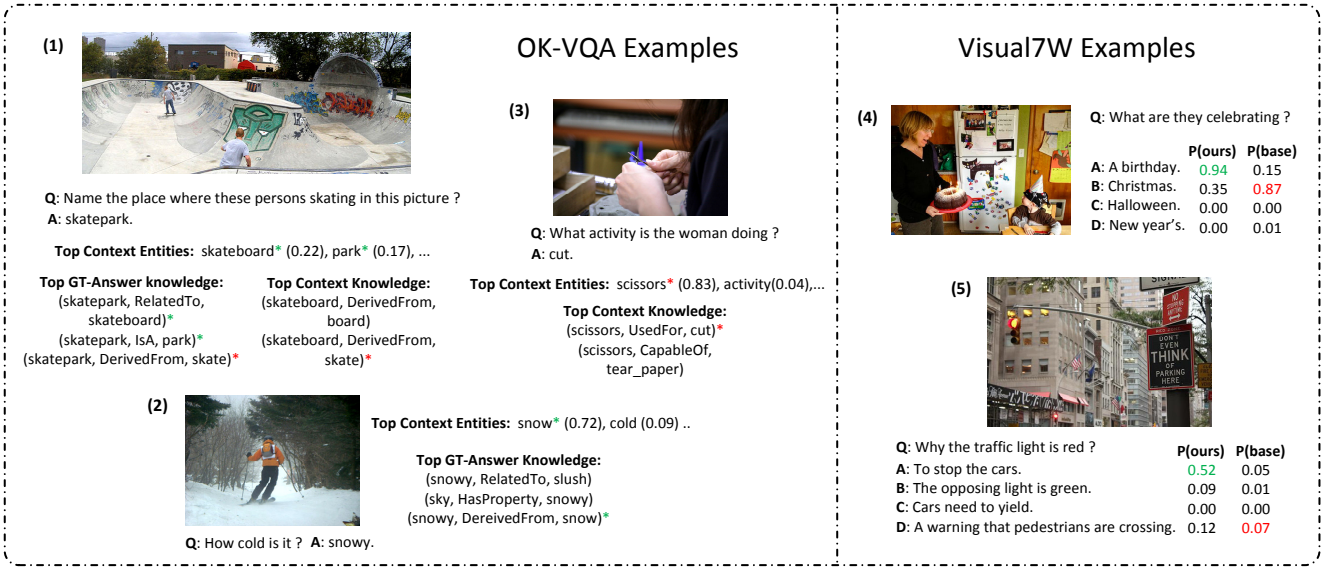

Figure 4: Visualized examples on the OK-VQA and Visual7W datasets. On the OK-VQA dataset, we provide the top-weighted context entities and the top-weighted knowledge that is aggregated into the ground-truth answer entity (*Top GT-Answer knowledge*) or the context entities (*Top Context Knowledge*). We mark the potentially useful entity connections with star symbols. As for the Visual7W dataset, we show the results of the *BAN+KG-Aug* model and compare with the *BAN* baseline model. Results show that external knowledge (e.g., *cake-RelatedTo-birthday* in example (4)) can help inferring the correct answer.

features contribute most is that they explicitly establish connections (graph reasoning paths) between question-image context and all candidate answers (i.e., $\boldsymbol{\beta} \in \mathcal{R}^{|\mathcal{A}|}$), which inherently encodes the external graph knowledge. Interestingly, using $\vec{\boldsymbol{e}}^{(\text{ctx})}$ alone (i.e., "w/o $\{\vec{\boldsymbol{u}}, \vec{\boldsymbol{e}}^{(\text{ans})}\}$") works even worse than doing nothing about question augmentation (1.19% > 0.91%). This fact further implies the difficulties of incorporating external knowledge and necessities of modeling the answer entities in a common knowledge space. Another observation is that the performance drops significantly when we disable the knowledge aggregation mechanism (we disable it by setting the aggregating gate $\gamma_e^{(t)}$ to zero).

Ablation Studies # 2: Effectiveness of Knowledge Sources. We incorporate external knowledge from two knowledge sources (ConceptNet and Wikidata) in our proposed method. Here we investigate the effectiveness of each knowledge source and the importance of using proper knowledge.

As shown in Table 4(b), removing the knowledge from Conceptnet leads to a performance loss of 0.38%, while for Wikidata knowledge, it is 0.22%. The results demonstrate that both the commonsense knowledge in ConceptNet and the factual knowledge in Wikidata are helpful for answering questions in the OK-VQA dataset.

Additionally, in order to validate the importance of using proper knowledge, we fabricate a "noisy knowledge source" by randomly permuting entities and relations in the original external knowledge graphs, where the connections between entities are noisy and unreliable. As expected, the noisy knowledge harms the model performance with an accuracy loss of 0.76%. Interestingly, this effect is similar to that of removing the question augmentation (i.e. "w/o $\{\vec{\boldsymbol{e}}^{(\text{ctx})}, \vec{\boldsymbol{u}}, \vec{\boldsymbol{e}}^{(\text{ans})}\}$"), implying that the unreliable knowledge prevent the model from exploiting the entities connections that are normally embedded in the auxiliary question features.

Visualizations. We visualize several examples in Fig. 4 to provide more insights. After inspecting lots of successful and failure cases on the OK-VQA dataset, we find that the model can leverage the connections between the context entities and candidate answer entities, which are built upon the external knowledge graphs. The context entities can be connected to the answer through various types of knowledge, such as *context entities knowledge* (e.g., Fig. 4 (3)) or *answer entities knowledge* (e.g., Fig. 4 (2)) or both (e.g., Fig. 4 (1)).

5 CONCLUSIONS

In this work, we propose the Knowledge Graph Augmented (KG-Aug) model, a practical and generic framework for incorporating large-scale external knowledge graphs into the VQA task. Our KG-Aug model learns to conduct context-aware knowledge aggregation on a retrieved knowledge subgraph and smoothly incorporates the aggregated knowledge into conventional VQA models to boost their performance. We validate the effectiveness of our model on various datasets, showing that the KG-Aug model exhibits superior performance and generalizes well under various experimental settings.

ACKNOWLEDGMENTS

This research is supported by National Natural Science Foundation of China Major Project No.U1611461 and National Key R&D Program of China under Grand No.2018AAA0102000.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998* (2017).

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

[3] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2612–2620.

[4] Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 740–750.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 363–370.

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[7] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2019. KnowIT VQA: Answering knowledge-based questions about videos. *arXiv preprint arXiv:1910.10706* (2019).

[8] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956* (2018).

[11] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*. 1564–1574.

[12] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325* (2016).

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[14] Guohao Li, Hang Su, and Wenwu Zhu. 2017. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *arXiv preprint arXiv:1712.00733* (2017).

[15] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*. 289–297.

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3195–3204.

[17] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*. 2654–2665.

[18] Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 451–468.

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[20] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8876–8884.

[21] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019. From Strings to Things: Knowledge-enabled VQA Model that can Read and Reason. In *Proceedings of the IEEE International Conference on Computer Vision*. 4602–4612.

[22] Robyn Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC*. 3679–3686. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf

[23] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.

[24] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2017. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. *arXiv preprint arXiv:1708.02711* (2017).

[25] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[26] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2018), 2413–2427.

[27] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).

[28] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630.

[29] Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving Question Answering over Incomplete KBs with Knowledge-Aware Reader. *arXiv preprint arXiv:1905.07098* (2019).

[30] Ziwei Zhang, Peng Cui, and Wenwu Zhu. 2018. Deep learning on graphs: A survey. *arXiv preprint arXiv:1812.04202* (2018).

[31] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4995–5004.