Find The Gap: Knowledge Base Reasoning For Visual Question Answering

Anonymous ACL submission

Abstract

We analyze knowledge-based visual question answering, for which given a question, the models need to ground it into the visual modality and retrieve the relevant knowledge from a given large knowledge base (KB) to be able to answer. Our analysis has two folds, one based on designing neural architectures and training them from scratch, and another based on large pre-trained language models (LLMs). Our research questions are: 1) Can we effectively augment models by explicit supervised retrieval of the relevant KB information to solve the KB-VQA problem? 2) How do task-specific and LLM-based models perform in the integration of visual and external knowledge, and multi-hop reasoning over both sources of information? 3) Is the implicit knowledge of LLMs sufficient for KB-VQA and to what extent it can replace the explicit KB? Our results demonstrate the positive impact of empowering task-specific and LLM models with supervised external and visual knowledge retrieval models. Our findings show that though LLMs are stronger in 1-hop reasoning, they suffer in 2-hop reasoning in comparison with our fine-tuned NN model even if the relevant information from both modalities is available to the model. Moreover, we observed that LLM models outperform the NN model for KB-related questions which confirms the effectiveness of implicit knowledge in LLMs however, they do not alleviate the need for external KB.

1 Introduction

Visual Question Answering (VQA) aims to answer a question in natural language form while visual information is needed to provide the correct answer. Though over the past years, the progress in this field has been remarkable, existing models still suffer in answering the questions requiring knowledge beyond the image. These models suffer from lacking extra knowledge integration or integrating noisy knowledge in their reasoning procedure even when the source of required knowledge is given. This problem is formulated as the so-called KB-VQA where in addition to the question and image, the source of external knowledge is given as input.

There have been various efforts to include external knowledge in the VQA task. These methods have two basic steps, retrieving related knowledge entries from the given knowledge bases, e.g., Wikipedia and ConceptNet (Liu and Singh, 2004), and predicting the final answer by integrating this extracted external information, and question, and visual data. However, these retrieval-based approaches suffer from two weaknesses: (i) They use some heuristics like keyword matching for retrieval, and there is no guarantee for retrieving semantically relevant knowledge and (ii) the knowledge graphs are very dense, and plenty of irrelevant knowledge is retrieved even if the required knowledge is retrieved. These irrelevant entries mislead the reasoning procedure. Since current information retrieval methods are based on simple heuristics and there is no guarantee for their retrieval accuracy, they hurt the final VQA results by retrieving irrelevant or noisy facts (Marino et al., 2021; Li and Moens, 2022).

With the introduction of the pre-trained large-scale language models and their vast power in solving various tasks, recent approaches aimed to use the implicit knowledge accumulated in these models, such as GPT-3 (Brown et al., 2020). These models (Hu et al., 2023; Shao et al., 2023; Yang et al., 2022) performed many efforts to use LLMs and guide commonsense reasoning in VQA models. Despite their remarkable results, these methods lack powerful vision processing and suffer from hallucination.

In this work, we propose to train a supervised knowledge retrieval model using the provided metadata in the KRVQA dataset (Cao et al., 2021). Our retrieval method uses contrastive loss to maximize the similarity of a question embedding with the em-

bedding of its supporting facts while minimizing its similarity to irrelevant facts. Moreover, we assume the corresponding scene graph (SG) of the image is given and utilize a similar retrieval module to extract the question-relevant visual information from the SG. In other words, we formulate both external knowledge and visual knowledge integration as retrieval tasks. Next, we integrate both external and visual retrieved knowledge to find the final answer through multiple hops of reasoning.

The retrieved knowledge from KG[1] and SG are integrated into both task-specific neural architectures (NN) and large language model(LLM) backbones. We analyze the strengths and weaknesses of the models regarding the question types and shed light on the remaining challenges in retrieval and integration steps for VQA task. In a nutshell, our contributions are as follows:

- We proposed a supervised model for retrieving relevant knowledge from the knowledge graph and scene graph conditioned on the question. Our results show the positive impact of empowering task-specific and LLM models with retrieved external and visual knowledge as well as their integration.

- We analyzed the impact of integrating external and visual knowledge in both task-specific neural architectures and LLMs for the VQA task. We found that LLM models outperform the NN model for KB-related questions which confirms the effectiveness of implicit knowledge in LLMs, however, they do not alleviate the need for external KB.

- We investigated the strength of the task-specific and LLM models in multi-hop reasoning and performing both visual and KB reasoning. Our results show that though LLMs are stronger in 1-hop reasoning, they suffer in 2-hop reasoning in comparison with our fine-tuned NN model even in the presence of the supporting knowledge from both modalities. This highlights the importance of using a strong module for integrating both modalities and performing complex reasoning on top of that.

[1]Since our external knowledge base is in graphical format, we use Knowledge Graph (KG) to mention our Knowledge Base (KB) interchangeably.

2 Related Works

We roughly divide recent VQA approaches into three groups as follows. The first group, plain VQA methods, focus on question and image integration approaches without including any extra information. The second group aims to feed extra information extracted from external knowledge bases in VQA models. In the last group of SOTA VQA models, we review methods that rely on the power of the pre-trained Large Language Models (LLMs) as implicit sources of knowledge for KB-VQA tasks.

2.1 Plain VQA methods

Similar to (Yu et al., 2018), MUTAN (Ben-Younes et al., 2017) aims to use bilinear models for merging visual and textual information in VQA tasks using a multimodal Tucker tensor decomposition (Tucker, 1966). MUTAN parametrizes bilinear interactions between visual and textual representations to learn high-level associations between question meaning and visual concepts in the image while avoiding suffering from the curse of dimensionality.

To focus on both the visual and textual context of questions, the authors of MCAN (Yu et al., 2019) aim to design a 'co-attention' model to associate keywords in questions with key objects in images. They propose a deep Modular Co-Attention Network (MCAN) that is made by in-depth cascading of Modular Co-Attention (MCA) layers. Each Modular Co-Attention layer (MCA layer) models the self-attention of questions and images, as well as the question-guided-attention of images jointly using a modular composition of two basic attention units.

The authors of (Ding et al., 2022) suggest extracting and using knowledge triplets from training data, instead of feeding external knowledge bases. They propose to extract and represent multimodal knowledge by explicit triplets during training of VQA task. They discuss the complementary of the extracted triplets to the existing knowledge graphs and leave combining external knowledge bases open for the future.

2.2 KB processing methods

The MuCKO model (Zhu et al., 2020) aims to answer the question by generating a semantic graph built on the dense captions of the image, a graph representing relevant knowledge triplets, and a spa-

tial visual graph where nodes represent region features. It performs question-guided inter and intra-graph attention to answer the question iteratively.

The authors of (Marino et al., 2021) use Multimodal-BERT (MMBERT) (Khare et al., 2021) for their multimodal embedding and use Relational Graph Convolutional Network (RGCN) (Schlichtkrull et al., 2018) to learn the representation of the concepts extracted from questions, Image, and KBs. To cover the knowledge required in OK-VQA, they construct a KB later. Their model can improve marginally while suffering from huge memory and time demands. Moreover, since the supporting KB and the OKVQA dataset have been collected separately, this model is not analyzable. There are remaining questions on how VQA models can benefit from external KB while avoiding upcoming noise.

MAVEx is proposed by (Wu et al., 2022) aiming to leverage and integrate knowledge from visual (Google image search), textual (Wikipedia), and commonsense knowledge sources (ConceptNet). Their goal is to validate a set of candidate answers using answer-based knowledge retrieval.

In (Guo et al., 2022), the authors offer Unifer, a unified end-to-end retriever-reader framework for KB-VQA. To reduce noisy retrieval from external KBs, they aim to evaluate retrieved triplets based on their effect on final accuracy. They pay a huge cost for generating pseudo-labels to evaluate all knowledge triplets for each question.

DMMGR (Li and Moens, 2022) represents the retrieved KB triplets in dynamic key-value memory format and converts visual content in a spatial graph representation.It performs a question-knowledge guided attention over the spatial visual graph to find the required visual information to answer the question.

2.3 Large Language Models as Implicit Knowledge Sources

Due to the revolutionary growth and power of large pre-trained models, the new trend in knowledge-based VQA aims to use large language-only pre-trained models like GPT-3 (Brown et al., 2020) as implicit sources of knowledge, instead of using external sources of knowledge. Most often, these methods replace the visual data (image) with its corresponding tags and/or captions and then solve an unimodal text-only problem.

The authors of (Gui et al., 2021) use a contrastive-learning-based module to retrieve knowledge from an external KB, and use GPT-3 to retrieve implicit knowledge by feeding image tags and captions as an input prompt. They integrate implicit and explicit knowledge and generate the final answer by integrating both implicit and explicit knowledge. The authors of (Yang et al., 2022) propose PICa which relies on GPT-3 in joint retrieval of relevant knowledge and reasoning. They verbalize the visual signal by replacing the image with its caption and/or tags to provide more in-context information for the language model.

The authors of (Lin et al., 2022) offer REVIVE targeting to integrate visual information, implicit knowledge, and explicit knowledge to generate the final answer. Similar to KAT, they have used a subset of Wikidata (Vrandečić and Krötzsch, 2014) as their KB. The authors of (Hu et al., 2023) propose Promptcap (Prompt-guided image Captioning) to control the visual entities in the generated caption using a textual query, and replace a general caption with a question-dependent caption.

To avoid losing visual information, the authors of (Salaberria et al., 2023) feed visual features alongside regional labels and captions to the transformer in their proposed Caption-Based Model(CBM). Prophet's (Shao et al., 2023) authors suggest including some candidate answers to enrich the prompt and improve final accuracy for LLM-based VQA.

It is worth mentioning that storing and retrieving knowledge implicitly in LLM parameters is prone to some errors happening during their knowledge storing and retrieving steps. These errors include but are not limited to 1. The knowledge hidden implicitly in LLM parameters is not necessarily complete, correct, or updated. 2. Correct memorization of knowledge during pre-training is not guaranteed, and 3. Retrieval of relevant knowledge is not guaranteed (He et al., 2022).

3 The Proposed Method

Our goal is to retrieve and integrate relevant knowledge triplets from the KGs to correctly answer a textual question related to a given image.

There are two main challenges in this task, first, retrieving relevant knowledge from the external knowledge base, and second, integrating retrieved knowledge, the question, and the image to generate the final answer. Regarding the first challenge, we propose a supervised model based on contrastive loss to maximize the similarity of the question em-

bedding with the embedding of its supporting facts while minimizing their similarity to irrelevant facts. Next, we replace our raw image with its corresponding scene graph and propose a similar contrastive-loss optimization method to retrieve the relevant visual information. We used the human-annotated scene graphs to represent the given visual data. For more information on the scene graph extraction and representation, one can check Appendix, section 6.1. Regarding integration and reasoning, we propose two frameworks based on task-specific neural architecture vs. using LLMs as the backbone to provide a comprehensive analysis of the impact of our retrieval method.

3.1 External Knowledge Retrieval

Since the most popular supporting knowledge bases provided for the VQA task are in graph format (Cao et al., 2021; Marino et al., 2021), we propose a knowledge retrieval method for graphical knowledge-bases (KGs) which represents each fact in a triplet form, (e_1, r_1, e_2) indicating first entity, relation, and second entity, respectively. As mentioned before, we propose an objective loss function to learn the representations of the triplets by contrasting positive and negative triplets. We convert each triplet (e_1, r_1, e_2) to a sentence $e_1 r e_2$. One can use transformation-based graph encoders such as TransE, QualE, etc (Rossi et al., 2021), or neural network-based models such as Relational Graph Convolutional Networks (RGCN) (Schlichtkrull et al., 2018) to initialize the triplet embeddings. However, the first group has been trained on a limited amount of KB data, and the RGCN family suffers from memory issues for data with a huge number of relations, and it only can learn embedding for entities and not for relations. Due to the availability of strong efficient pre-trained text-encoders, we used Roberta (Liu et al., 2019) for both plain text and triplet embedding initialization. We have used a feed-forward sub-layer on top of the Roberta representations to transform the embeddings to the same space. The architecture for our retrieval model is shown in Fig 1.

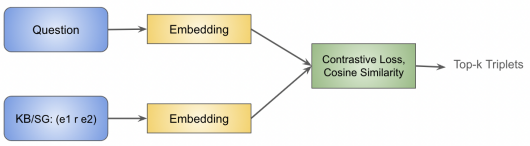

Figure 1: Architecture of our supervised retrieval model.

We use the following contrastive loss formulation,

$$L_{Ret}(q, KB) = \sum_{t^+ \in pos} \sum_{t^- \in neg} [d(q, t^+) - d(q, t^-)] \quad (1)$$

, where q is the question, KB is the set of all KB triplets, t^+ is a set of triplets appearing in the supporting reason for the given question, and t^- is a subset of all non-related triplets. $d(.,.)$ can be any distance measure, for which we use negative cosine similarity.

We use the same initialization method and architecture for both visual triplet retrieval from the scene-graph, and knowledge triplet retrieval from the external knowledge graph. It is worth mentioning that we have utilized the supportive reasons provided by KRVQA dataset for each question for supervising our retrieval module. We have used the image-related supporting facts (triplets) for supervising visual-retrieval training, and KB-related triplets for supervising KB retrieval training. an example of KRVQA dataset is shown in Fig 2. Since,

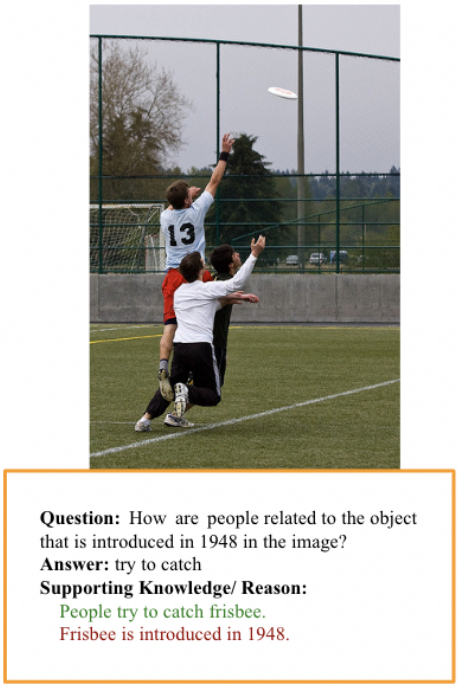

Figure 2: An example krvqa dataset. given image, question, answer, and Supporting reason. The green part of the reason is extracted from the image, and the red part from the external KB.

the external KB includes 2339 relations, 102343 entities, and 225434 facts, we first use keyword

matching to find a smaller related subgraph before passing the graph to the CL-based retrieval step. We have extracted tags from the image, and keywords from the question, and extracted a subgraph by traversing two hops from each node, after finding KB nodes with matching keywords. It is worth mentioning that, this small subgraph is still very huge and has an order of $10K$ edges.

3.2 Integration and Reasoning Using Task-specific Neural Architectures

In our proposed task-specific architecture, we do multi-hop reasoning by integrating the retrieved KG and SG triplets in an iterative manner using cross-attention units and generate the final answer as follows.

$$query_0 = question, \tag{2}$$
$$query_t = Att(query_{t-1}, KB_{ret}) + Att(query_{t-1}, SG_{ret})$$

In other words, to handle several hops of reasoning, we iteratively update the cross-attention query (initialized by question embedding) by attending to retrieved KB and SG triplets. In the end, we combine the output of 2 cross-attention modules to perform the final VQA classification. The complete architecture of our task-specific neural model is shown in Fig 3.

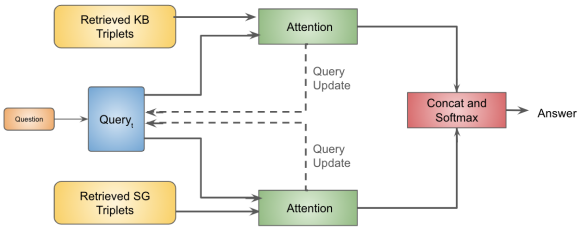

Figure 3: Architecture of our proposed task-specific neural architecture. The attention modules iteratively update a query to solve multi-hop reasoning. The integration layer combines external and visual knowledge to provide the final output using a classifier layer. We initialize the query with question embedding and use KB and SG triplets as keys and values. the dashed arrows update the query iteratively using the output of attention-modules.

3.3 Integration and Reasoning using Large Language Models

Similar to (Hu et al., 2023; Shao et al., 2023; Yang et al., 2022), we use a combination of the question and image caption as the input prompt for our LLM building block. For each test example, we find 32 training examples using maximum cosine similarity of image+question embeddings to form our few-shot prompts. We use CLIP(Radford et al., 2021) to embed the questions and images. To feed the supporting knowledge, we convert each triplet (e_1, r_1, e_2) to a sentence $e_1 r e_2$, and inject it into the prompt as can be seen in the knowledge section as in Figure 4. For the sake of fair comparison, we do not feed extra information such as image tags or candidate answers in our prompts. We use Promptcap captioner (Hu et al., 2023) to convert our images to text, and add these captions to the context section of the prompts. Due to its significant accuracy of "davinci-002" for in-context learning, like (Hu et al., 2023; Shao et al., 2023) we use it as our LLM engine.

4 Experimental Results

In the following sections, we will provide experimental results to show the impact of supervised knowledge retrieval and integration in both task-specific neural architectures and LLM frameworks for the VQA task. We aim to design experiments and provide results to answer the following questions: 1- Can we effectively augment task-specific and LLM VQA models by supervised retrieval of the relevant external and visual information? 2- How do task-specific and LLM-based models perform in multi-hop reasoning and integration of both external and visual sources of information? 3. To what extent the implicit knowledge of LLMs can cover the required external knowledge for VQA task?

4.1 Datasets

Cao et al. propose knowledge-routed visual question reasoning (KRVQA) dataset that aims to prevent shortcut learning by generating non-crowdsourced question–answer pairs with their supporting reasons indicating which part of the reason is from visual, and which part is from external knowledge. It provides meta-data to analyze the strengths and weaknesses of the VQA models, for example, question type, supporting reasons divided into KB and visual parts, and number of reasoning hops to answer the questions.

KRVQA is to date the largest knowledge-based VQA dataset. It contains 32910 images and 157201 QA pairs. They are split into the train, validation, and test set at the ratio of 60%, 20%, and 20%,

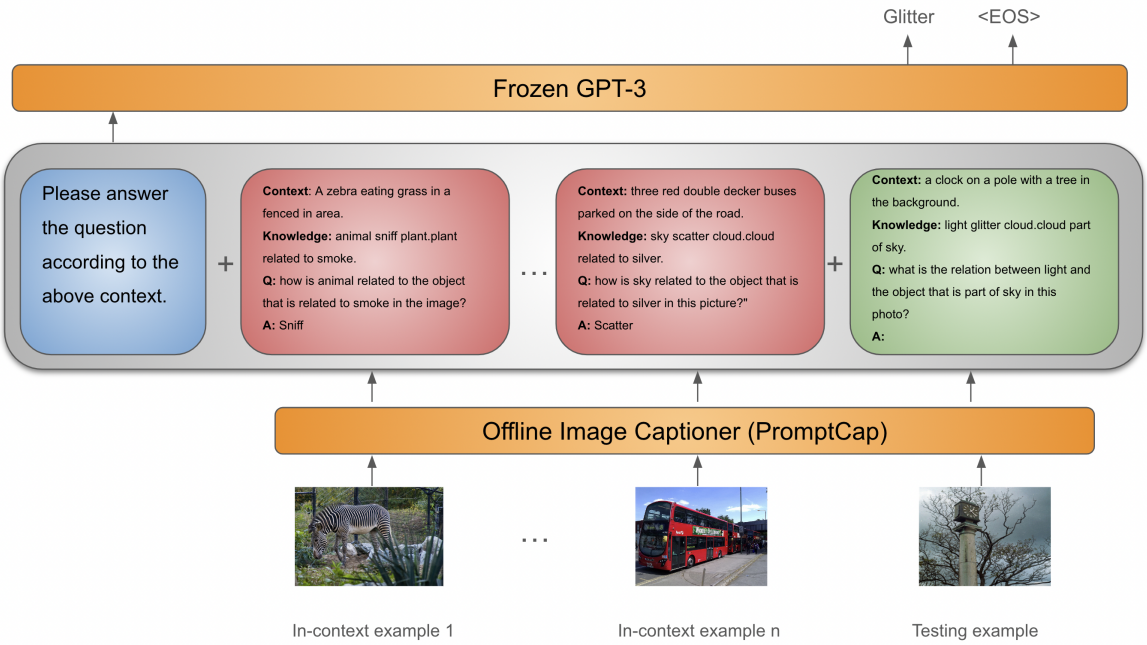

Figure 4: Architecture of our LLM model. The relevant knowledge is given as a part of the prompt

respectively. The questions in this dataset use external knowledge from FVQA dataset 225434 factual triplets (Wang et al., 2017). This external knowledge is being extracted from DBpedia(Auer et al., 2007), ConceptNet(Speer et al., 2017), and WebChild(Tandon et al., 2014). One of the current popular VQA datasets is OK-VQA (Marino et al., 2019). This dataset does not provide supporting facts/reasons for each question and lacks supporting metadata for our analysis goal. In this work, we use KRVQA dataset as its supporting meta-data helps to train our retrieval modules and provide a detailed analysis.

4.2 Implementation Details

We train our model with PyTorch (Paszke et al., 2019) for all experiments using one GPU with 49GB of memory. For the contrastive loss, we treated all the other samples in a batch as negative samples. Our model is trained by AdamW (Loshchilov and Hutter, 2018) optimizer with 200 epochs, where the batch size is 256 and the learning rate is set to $1 * 10^{-4}$. Cascading 4 layers of query-update and reasoning provides the best accuracy for integration and reasoning. We use top-1 accuracy as in (Cao et al., 2021) for a fair comparison.

4.3 Retrieval Result

The accuracy of our retrieval module for both scene-graph and knowledge base is shown in Table 1. While top-1 retrieval accuracy for both SG and KB is around 60%, top-100 accuracy for both models is almost 100%. We can see that top-1 accuracy without training (using cosine similarity of text embedding) is less than half of the accuracy for the supervised retrieval.

	KB Retrieval	SG Retrieval
Supervised	59.10	60.00
unsupervised	17.5	27.12

Table 1: Top-1 accuracy for our Retrieval on KRVQA dataset.

4.4 Result and Discussion

Table 2 shows accuracy for VQA task using KRVQA dataset. **NN+**(KB_{ret}+SG_{ret}) is our task-specific neural architecture while our retrieval model is used to retrieve the relevant part of the SG and KG (given the ground-truth scene graph and a relevant subgraph of the KG with preprocessing), as explained in section 3.2. **LLM+**(KB_{ret}+SG_{ret}) is our model with LLM backbone, enriched with relevant KB and SG triplets extracted using the method introduced in section 3.1.

Our enriched LLM model improves accuracy (40.5%) in comparison with Promptcap(Hu et al., 2023) (19.8%) which is one of the most recent VQA frameworks using LLM.

Our NN model has made more than 13% improvements in comparison with the best NN model,

DMMGR, which uses key-value memory networks to combine external knowledge in VQA tasks. This result is not a fair comparison though since we provide the ground-truth SG but our goal is to show the impact of supervised knowledge retrieval conditioned on access to the complete SG and KB. This result demonstrates the positive impact of training a supervised knowledge retrieval module for both external and visual knowledge besides a strong reasoning module to avoid noise propagation and enrich the model to perform better reasoning. The results show that enriching a base LLM model (Promptcap) by adding retrieved knowledge makes a significant improvement in accuracy, though still, a fine-tuned reasoning module leads to higher accuracy than LLM-based model.

Q-type (Cao et al., 2021)	8.12
LSTM (Cao et al., 2021)	8.81
FiLM (Perez et al., 2018)	16.89
MFH (Yu et al., 2018)	19.55
UpDown (Anderson et al., 2018)	21.85
MCAN (Yu et al., 2019)	22.23
Mucko (Zhu et al., 2020)	24.00
KM-net (Cao et al., 2019)	25.19
MUKEA (Ding et al., 2022)	27.38
DMMGR (Li and Moens, 2022)	31.8
Promptcap (Hu et al., 2023)	19.8
NN+(KB_{ret})	22.28
NN+(SG_{ret})	38.31
NN+(KB_{ret}+SG_{ret})	44.36
LLM+(KB_{ret})	21.6
LLM+(SG_{ret})	37.4
LLM+(KB_{ret}+SG_{ret})	40.50

Table 2: Top-1 accuracy percentage comparisons among different models on the KRVQR dataset.

As discussed in the appendix, we tried to replace the human-annotated scene graph with an approximated scene graph, but due to the low accuracy of current scene graph approximation methods, it did not lead to notable improvement and interpretable results. The remarkable improvement made by adding retrieved scene graph triplets demonstrates the importance of representing visual data in a high-level format instead of a low-level and abstract feature format.

4.5 Question-type Analysis

The KRVQA dataset is generated using 7 different question-answer templates as shown in Table 3. We provided the accuracy of both our NN, and LLM-based models for each question type. $+KB_{ret}$ and $+SG_{ret}$ indicate feeding retrieved external and visual knowledge in the model respectively, while $+KB_{GT}$ and $+SG_{GT}$ indicate injecting ground-truth triplets in the models to find the upper-bound accuracy using our reasoning module.

From results in Table 4, we can see that the proposed reasoning model is strong enough to find the final answer by combining the supporting facts, although even in the presence of ground-truth supporting facts/reasons, questions type 2, 3, and 5 are harder to answer using our NN model. Moreover, by comparing the accuracy using the retrieval model versus the upper limit of accuracy using ground truth triplets, our retrieval methods hurt more for question types 2,3, and 5. We conjecture that retrieving relevant facts is harder when the head entity is missing. As shown in Table 5, for our retrieval-based model, KB-related questions are harder than non-KB-related models, although our retrieval accuracy for both SG and KB is similar. Moreover, the NN model has similar accuracy for 1-hop and 2-hop questions.

Results show that question types 3 and 5 are harder for the LLM model even in the presence of ground-truth knowledge. Since for question types 3 and 5, the answer belongs to the second hop, it might make reasoning more complicated and prone to noise. Another interesting result is that almost for all one-hop questions LLM beats the NN model, though for 2-hop questions it performs worse than NN which confirms the weakness of LLM models in multi-hop reasoning. In Table 5, we can see that the LLM models work better than their corresponding NN models for KB-related questions which confirms the partial existence of implicit knowledge in LLMs.

4.6 Ablation Study

We analyzed the impact of each source of knowledge (visual vs external) for our VQA task in Table 6. Moreover, We reported the effect of injecting retrieved knowledge vs the ground-truth knowledge which indicates the upper-bound for our models. The NN-None model only integrates question representation with image CLIP representation, while the LLM-None model does not use any knowledge in the prompts.

NN model: The results show that both proposed KB and SG retrieval models make improvements in accuracy. Moreover, we can see that our proposed multi-hop reasoning is strong enough to make reasoning and find the final answer from the ground-truth supporting facts (accuracy equal to 89.48%). The huge gap between the None and SG_{ret} model

Qtype	Question Semantics	Answer	Reason
0	What is the relation of $< A >$ and $< B >$?	$< R >$	(A, R, B)
1	What is $< A > < R >$?	$< B >$	
2	What $< R > < B >$?	$< A >$	
3	What is the relation of the object that $< A > < R_1 >$ and $< C >$?	$< R_2 >$	$(A, R_1, B), (B, R_2, C)$
4	What is the relation of <A> and the object that $< R_2 > < C >$?	$< R_1 >$	
5	What $< A > < R_1 > < R_2 >$?	$< C >$	
6	What $< R_1 > < R_2 > < C >$?	$< A >$	

Table 3: KRVQA templates for the question-answer generation(Cao et al., 2021)

Q-type	0	1	2	3	4	5	6	Overall
NN+(KB_{ret}+SG_{ret})	63.34	56.76	23.63	14.75	54.55	10.28	54.29	89.31
NN+(KB_{GT}+SG_{GT})	99.74	92.39	86.82	87.44	95.96	74.24	90.86	44.36
LLM+(KB_{ret}+SG_{ret})	60.71	57.31	33.95	12.58	51.93	14.49	51.78	40.5
LLM+(KB_{GT}+SG_{GT})	99.01	96.00	98.52	74.47	92.56	74.54	88.24	88.90

Table 4: Accuracy for each question type.

Q-type	KB-related	No-KB-related	1-hop	2-hop
NN+(KB_{ret}+SG_{ret})	34.63	43.33	37.99	38.83
NN+(KB_{GT}+SG_{GT})	85.57	94.25	90.46	88.57
LLM+(KB_{ret}+SG_{ret})	39.36	41.59	43.82	37.67
LLM+(KB_{GT}+SG_{GT})	91.19	91.15	98.09	80.68

Table 5: Accuracy Analysis.

additional information	None	KB_{ret}	SG_{ret}	KB_{ret}+SG_{ret}	KB_{GT}	SG_{GT}	KB_{GT}+SG_{GT}
NN	15.36	22.28	38.31	44.36	38.98	67.06	89.48
LLM	19.8	21.6	37.4	40.5	36.60	58.60	87.90

Table 6: Ablation Study

shows the importance of providing a more high-level (object-level) representation of the visual data, despite a low-level abstract representation of an image or a general caption.

LLM model: The results for LLM-based model support the previous findings, though similar to the results in the previous section (Table 2), we can see that LLM models are moderately less accurate in comparison with the task-specific fine-tuned model (NN), which highlights the impact of a strong knowledge integration and reasoning module. Supporting the findings for NN model, the huge improvement resulting from adding visual knowledge retrieval demonstrates that a caption does not adequately represent the visual data, and we need an alternative approach to feed the visual data.

5 Conclusion and Future Work

We proposed a retrieval and reasoning framework for knowledge-based visual question answering. We trained a contrastive-loss model to retrieve supporting facts from external knowledge as well as from the scene graph. We used the KRVQA dataset, the largest existing VQA dataset, to train both retriever and reasoning modules.

Our results show the positive impact of empowering task-specific and LLM models with retrieved external and visual knowledge. We investigated the strength of the task-specific and LLM models in multi-hop reasoning, as well as integrating visual and KB reasoning. Our results show that though LLMs are stronger than the fine-tuned NN model in 1-hop reasoning, they suffer in 2-hop reasoning even in the presence of the supporting knowledge from both modalities. This highlights the importance of using a strong module for the integration of both sources of information and performing complex reasoning. We found that LLM models outperform the NN model for KB-related questions which confirms the effectiveness of implicit knowledge in LLMs, however, they do not alleviate the need for external KB.

6 Limitations

- We trained and executed retrieval and reasoning parts separately. We will investigate the impact of dynamic fact retrieval interchained with reasoning in the future.

- Due to the huge size of our external knowledge base, we retrieve the relevant knowledge for each question only one time before starting the reasoning step. we will replace this static retrieval module with a dynamic retrieval module which refines the retrieved knowledge after each step of reasoning.
- we used human-annotated scene graphs provided in the Visual Genome dataset, as approximated scene graphs do not generate acceptable accuracy. In the future, we have to replace VG scene graphs with another automatically generated high-level representation of the image.
- Some commonly-used VQA datasets like OK-VQA do not provide supporting facts/reasons for each question and lack supporting metadata to analyze the strengths and weaknesses of the reasoning models. Therefore, we did not provide results on this dataset. In the future, we try to modify our models and perform experimental results on more datasets.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

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

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*, pages 2612–2620.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qingxing Cao, Bailin Li, Xiaodan Liang, and Liang Lin. 2019. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *arXiv preprint arXiv:1909.10128*.

Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2758–2767.

Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. 2022. A unified end-to-end retriever-reader framework for knowledge-based vqa. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2061–2069.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975.

Yash Khare, U Priyakumar, VIRAJ BAGAL, CV Jawahar, Adithi Devi, and Minesh Mathew. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa.

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*, 123:32–73.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Mingxiao Li and Marie-Francine Moens. 2022. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10983–10992.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *arXiv preprint arXiv:2206.01201*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. 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.

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/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–49.

Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. 2023. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*, 212:118669.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 523–532.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Ledyard R Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2712–2721.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. 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.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959.

Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. 2023. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2915–2924.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*.

Appendix

6.1 Scene Graph Computation

We treated both the knowledge base and visual signal in graph format. We proposed a supervised retrieval method to retrieve relevant triplets from the external knowledge graph and scene graph. Since the images in the KRVQA dataset are taken from the Visual Genome dataset (Krishna et al., 2017), we have used the scene graphs provided by the VG dataset in our method. These scene graphs on average include 36 objects and 22 relationships per image.

To have a more realistic setting, we made efforts to provide results with approximated scene graphs as well. Given an image I, we used Faster-RCNN (Ren et al., 2015) to identify a set of objects $O = \{o_i\}_{i=1}^{K}, (K = 36)$, where each object o_i is associated with a visual feature vector $v_i \in R^{d_v}, (d_v = 2048)$, and a corresponding label. The final scene graph is shown as $G_V = (V^V, E^V)$ over objects O, where $V^V = v_i^V{}_{i=1}^{K}$ is the node set and each node v_i^V corresponds to a detected object o_i. We have trained our scene graph generation method using the training data introduced in (Xu et al., 2017) (VG150). VG150 includes the most frequent 150 object categories and 50 predicates of the VG dataset. We use 70% of the images for training and the remaining 30% for testing. We generated scene graphs using both VS3 (Zhang et al., 2023) which used GLIP (Li et al., 2022) as a building block for object and feature recognition, as well as the scene graph benchmark proposed by (Tang et al., 2020). Unfortunately, scene graph detection quality is low, and error propagation from scene graph generation hampers the retrieval accuracy from 60% to around 19% (as shown in table 1). Therefore, we use Visual Genome provided scene graphs in our work to avoid error propagation from the scene graph to the next steps of the model and have a better analysis of the strengths and limitations of our retrieval and reasoning model.