Bridging the Gap: Advancing Commonsense Question Answering with Integrated Multi-Modal Knowledge

Yongqiang Zhao (yongqiangzhao@stu.pku.edu.cn) and Zhi Jin (zhijin@pku.edu.cn)
School of Computer Science, Peking University
5 Yiheyuan Road, Haidian District, Beijing, 100871, China

Feng Zhang (zhangfeng@nudt.edu.cn), Xinhai Xu (xuxinhai@nudt.edu.cn), and Donghong Liu (2994665156@qq.com)
Academy of Military Sciences
128 Xiangshan Road, Qinglongqiao Street, Haidian District, Beijing, 100089, China

Abstract

Most current research on commonsense question answering (CQA) has focused on proposing different techniques in natural language processing and text information retrieval. However, for human cognition, retrieving and organizing desired answers from text knowledge related to commonsense questions is far less intuitive and comprehensive than it is when using multi-modal knowledge, such as related images and videos. Motivated by this, we propose a framework for trying the acquisition of diverse modal information, and embedding and integrating it into CQA tasks, further improving the performance and user experience. Specifically, this paper proposes the integration of multi-modal knowledge, including images, image description statements, image scene graphs, and knowledge sub-graphs, into a CQA system. It introduces a parallel embedding technique for this multi-modal knowledge and employs an alignment-interaction-fusion mechanism to facilitate the seamless integration of this multi-modal knowledge. Through extensive experiments, the effectiveness and superiority of our proposed method are demonstrated.

Keywords: commonsense question answering framework; multi-modal knowledge acquisition; parallel embedding; alignment-interaction-fusion

Introduction

The commonsense question answering (CQA) system is a type of conversational system that aims to analyze and comprehend user-posed commonsense questions and provide corresponding answers or solutions (Palta & Rudinger, 2023). CQA systems have been widely applied in various scenarios, such as online customer services, intelligent assistants, and search engines (He, Gutiérrez-Basulto, Pan, et al., 2023; Qin et al., 2023; Lan et al., 2022; Zhu et al., 2022).

Existing studies (Zou, Zhang, & Zhao, 2023; Khashabi et al., 2020; Dou & Peng, 2022; M. Zhang, He, & Dong, 2024) have concentrated their efforts on natural language processing and knowledge base integration via the information retrieval paradigm. Among them, the utilized knowledge mainly includes textual modal data, such as free text knowledge and structured knowledge. However, for human cognition, retrieving and organizing answers from text knowledge related to commonsense questions is far less intuitive and comprehensive than doing so from multi-modal knowledge such as related images and videos. For example, when answering questions about the spatial relationships between objects, or about the color, shape, and size of objects, information from other modalities can be much more helpful. As illustrated in Figure 1, questions can be answered well by information from corresponding images. Inspired by this, we aim to leverage multi-modal knowledge to enhance CQA systems.

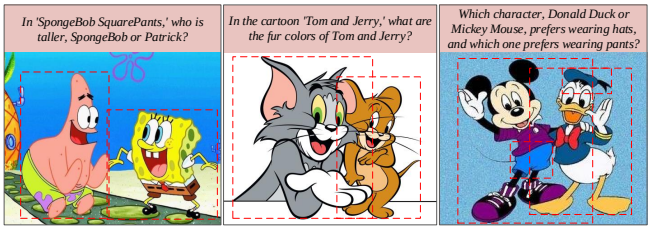

Figure 1: Multi-modal knowledge provides auxiliary information for the commonsense question-answering task.

Specifically, this paper proposes an innovative method for acquiring multi-modal knowledge by utilizing advanced techniques such as keyword retrieval (Campos et al., 2018), image search (M. Wang, Wang, Qi, & Zheng, 2020), image description generation (Nguyen, Suganuma, & Okatani, 2022), and image scene graph generation (J. Yang et al., 2022). To achieve effective feature representation for different types of data, this paper introduces the parallel embedding mechanism for acquiring multi-modal knowledge, which consists of three types of encoders: language transformer encoder, vision transformer encoder, and graph neural network. Finally, an alignment-interaction-fusion mechanism for multi-modal embedded features is specifically designed and constructed to better fuse embedded features from different modalities and enhance the model's overall generalization ability.

Our contributions can be summarized as follows:

(1) A novel CQA framework based on the fusion of multi-modal knowledge, which combines textual information with multi-modal knowledge information to obtain accurate question answers that better meet user requirements.

(2) A method for acquiring multi-modal knowledge based on keyword retrieval, image search, image captioning, and image scene graph generation. Using these techniques, we have built a novel multi-modal resource base (MRB).

(3) A method for parallel embedding of multi-modal knowledge, along with a mechanism for aligning-interacting-fusing them, aiming to achieve diverse feature representations and compatibility with different types of data, thereby effectively integrating multi-modal information.

(4) Extensive experiments are conducted to demonstrate

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

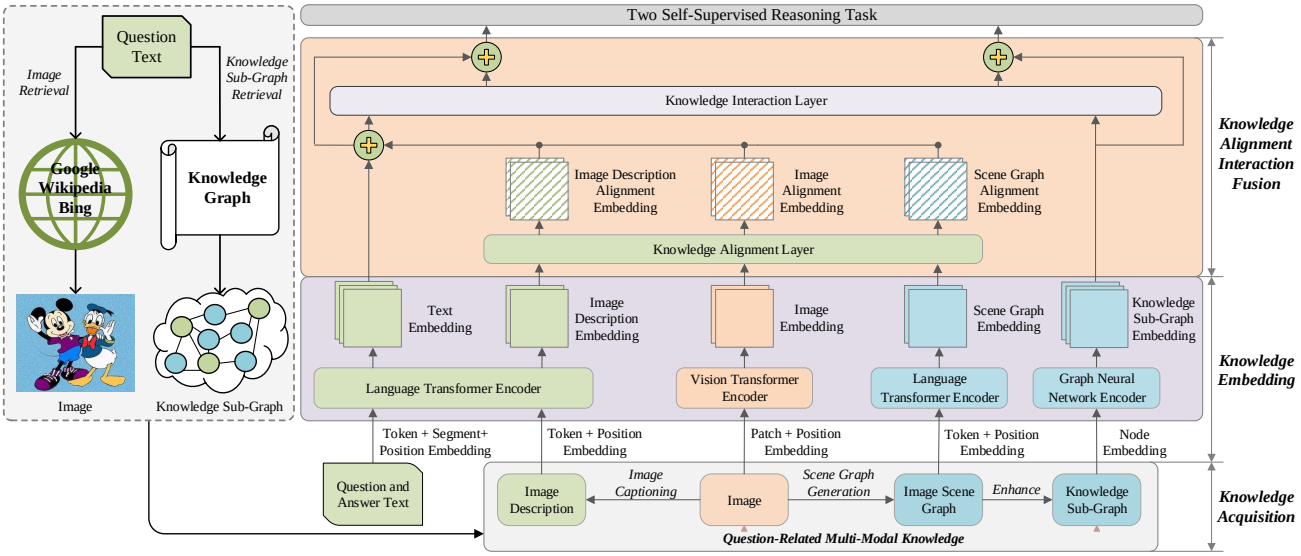

Figure 2: An overview of our multi-modal knowledge enhancement framework (MKE).

the effectiveness and superiority of the proposed method, achieving an accuracy improvement of 4.0% and 3.4% on the OBQA and CSQA datasets, respectively.

Related Work

Auxiliary knowledge augmentation is an active research field aimed at improving the performance of CQA systems (Q. Chen et al., 2023). Existing CQA systems mainly utilize free text knowledge and structured knowledge as auxiliary knowledge. **Free text knowledge**, as a fundamental form of auxiliary knowledge, is primarily extracted from large-scale textual corpora (Rosset et al., 2020). This type of knowledge encompasses domain-specific knowledge, commonsense knowledge, and textual knowledge obtained from other sources, such as social media, etc. Existing work (Guu, Lee, Tung, Pasupat, & Chang, 2020; Borgeaud et al., 2022) has primarily focused on utilizing various methods to extract free textual knowledge from large-scale corpora and use them effectively. The vast amount of user-generated content on social media platforms contains valuable knowledge, such as online encyclopedias, user comments, posts, tweets, and more. CQA systems leverage social media knowledge (Rogers, Gardner, & Augenstein, 2023; G. Xu et al., 2021; Li et al., 2020; R. Sun et al., 2020; Cui, Lan, Pang, Guo, & Cheng, 2020; H.-Y. Yang & Silberer, 2022) to acquire real-time information and user perspectives, enabling better answers to questions related to social media. **Knowledge graphs**, as a structured representation of knowledge, have been widely applied in CQA systems (Y. Sun, Shi, Qi, & Zhang, 2022; Y. Xu et al., 2021; Pan et al., 2023). Researchers have continuously explored how to better construct and utilize knowledge graphs to enhance the performance of CQA systems (Tian, Jing, He, & Liu, 2021; S. Liu, Chang, Liang, Chakraborty, & Driggs-Campbell, 2021).

Method

This paper proposes a multi-modal knowledge enhancement framework (MKE) for CQA, which integrates knowledge of multi-modalities with the question-answer textual information to more effectively meet user needs. The core design of this framework includes three parts, as shown in Figure 2.

Knowledge Acquisition: Acquire various knowledge from multiple modalities relevant to the question, including images, textual descriptions of images, scene graphs of images, and knowledge sub-graphs. This enables the provision of rich and diverse commonsense knowledge.

Knowledge Embedding: Use multiple parallel encoders to embed the input question-answer text and knowledge from various modalities acquired earlier, capturing the corresponding features of information from these modalities.

Knowledge Alignment, Interaction, and Fusion: Effectively leverage the acquired information from multiple modalities by aligning, interacting, and fusing the features of the information from different modalities.

Knowledge Acquisition

The existing methods for multi-modal knowledge acquisition in the context of images primarily rely on utilizing only the respective images. However, such approaches may not be able to capture deeper relationships between objects, as well as object properties such as the object's color, shape, size, and so on. To address this issue, a new multi-modal knowledge acquisition method is proposed. This method integrates keyword-based extraction, image search, image description generation, and scene graph generation. This enables a more comprehensive and accurate extraction of fine-grained visual information such as entities, entity relationships, and entity attributes contained in the image. Specifically, we first utilize the keyword extraction algorithm YAKE (Campos et al.,

2018) to extract keywords from the question:

$$YAKE(Question) = \{(k_i, s_i)\}_{i=1}^{K} \quad (1)$$

where k_i represents keyword, s_i represents the keyword score, and K is the number of keywords. Then, we use these pairs of keywords to retrieve relevant images from image search engines (Google, Wikipedia, and Bing) (Y. Chen et al., 2023). Subsequently, the image captioning and scene graph generation algorithms (Nguyen et al., 2022; J. Yang et al., 2022) are employed to process the retrieved images and obtain corresponding detailed image descriptions and scene graphs:

$$Caption(Image) = \{d_i\}_{i=1}^{M} \quad (2)$$

$$SceneGraph(Image) = \{(h_i, r_i, t_j)\}_{i=1}^{N} \quad (3)$$

where d_i represents the i-th word in the description, and M represents the number of words. (h_i, r_i, t_i) represents a triplet in the scene graph, N represents the number of triplets. At the same time, inspired by existing work (Yasunaga et al., 2022), for each question, we also retrieve the knowledge sub-graph from ConceptNet (Speer, Chin, & Havasi, 2017).

Knowledge Embedding

To achieve diverse feature representation and compatibility with different types of information, we propose a multi-channel parallel embedding method. Specifically, language transformer encoders (LTE) are employed for question-answer text, image description, and scene graph, while a vision transformer encoder (VTE) is used for image. Furthermore, the knowledge sub-graph, enhanced by the image scene graph, is embedded using a graph neural network (GNN).

Text Embedding. For a given token sequence of question-answer text $T = \{t_1, ..., t_I\}$. We prepend t_{int} to the original question-answer text T, and the input representation is calculated by summing the token embeddings, segment embeddings, and position embeddings for each token:$\{t_{int}^{(0)}, t_1^{(0)}, ..., t_I^{(0)}\}$. Then, we compute the output representation for each layer to obtain the embedding E_{Text}:

$$\{t_{int}^{(w)}, t_1^{(w)}, ..., t_I^{(w)}\} = \text{LTE}(\{t_{int}^{(w-1)}, t_1^{(w-1)}, ..., t_I^{(w-1)}\}) \quad (4)$$

where w represents the layer index in the LTE, and the specific LTE used is RoBERTa-Large (Y. Liu et al., 2019).

For a given token sequence of image description $\{d_1, ..., d_M\}$ and scene graph $\{h_1, r_1, t_1, ..., h_N, r_N, t_N\}$, the input representation is calculated by summing the token embeddings and position embeddings for each token: $\{d_1^{(0)}, ..., d_M^{(0)}\}$, $\{h_1^{(0)}, r_1^{(0)}, t_1^{(0)}, ..., h_N^{(0)}, r_N^{(0)}, t_N^{(0)}\}$. Then, we compute the output representation for each layer to obtain the embedding $E_{Caption}$ and embedding $E_{SceneGraph}$:

$$\{d_1^{(x)}, ..., d_M^{(x)}\} = \text{LTE}(\{d_1^{(x-1)}, ..., d_M^{(x-1)}\}) \quad (5)$$

$$\{h_1^{(y)}, r_1^{(y)}, t_1^{(y)}, ..., h_N^{(y)}, r_N^{(y)}, t_N^{(y)}\} = \text{LTE}(\{h_1^{(y-1)}, r_1^{(y-1)}, t_1^{(y-1)}, ..., h_N^{(y-1)}, r_N^{(y-1)}, t_N^{(y-1)}\}) \quad (6)$$

where x and y represent the layer index in the LTE, and the specific LTE used is pretrained RoBERTa-Base.

Image Embedding. A given image is divided into a sequence of patches $\{p_1, ..., p_O\}$, where O denotes the number of patches. The input representation is calculated by summing the patch and position embeddings for each patch:$\{p_1^{(0)}, ..., p_O^{(0)}\}$. Then, the output representation for each layer is computed to obtain the embedding E_{Image}:

$$\{p_1^{(z)}, ..., p_O^{(z)}\} = \text{VTE}(\{p_1^{(z-1)}, ..., p_O^{(z-1)}\}) \quad (7)$$

where z is the VTE (Dosovitskiy et al., 2021) layer index.

Graph Embedding. We use the VCU method (McInnes, 2016) to combine the scene graph extracted from the image with the retrieved knowledge sub-graph. This leverages fine-grained entity and relationship information in the image to obtain the enhanced knowledge sub-graph $\{g_1, ..., g_F\}$, where F represents the number of nodes. Then, we connect g_{int} to the entity-linked nodes in the enhanced knowledge sub-graph, resulting in the final knowledge sub-graph $\{g_{int}, g_1, ..., g_F\}$. Subsequently, we compute the output representation for each layer to obtain the embedding E_G:

$$\{g_{int}^{(l)}, g_1^{(l)}, ..., g_F^{(l)}\} = \text{GNN}(\{g_{int}^{(l-1)}, g_1^{(l-1)}, ..., g_F^{(l-1)}\}) \quad (8)$$

where l is the GNN (Velickovic et al., 2017) layer index.

Knowledge Alignment, Interaction, and Fusion

Knowledge Alignment Layer Multi-modal knowledge is heterogeneous at low-level representation but unified at high-level semantics (Cao et al., 2022). The alignment layer focuses on aligning different modalities of knowledge, ensuring their compatibility, and enabling meaningful cross-modal interactions in a shared feature space. Since both image descriptions and scene graphs are derived from images, after obtaining their embeddings, this paper utilizes linear transformations to align their dimensions (Eq. 9) and then combines the aligned embeddings of different modalities with the question-answer text embeddings (Eq. 10).

$$\bar{E}_{Image}, \bar{E}_{Caption}, \bar{E}_{SceneGraph} = \text{Linears}(E_{Image}, E_{Caption}, E_{SceneGraph}) \quad (9)$$

$$\hat{E}_T = \text{Add}(E_{Text}, \bar{E}_{Image}, \bar{E}_{Caption}, \bar{E}_{SceneGraph}) \quad (10)$$

where *Linears* are linear dimension transformations, *Add* denotes feature concatenation, $\bar{E}$ represents the aligned embedding features, and $\hat{E}_T$ represents the question-answer text embedding after incorporating the aligned embeddings.

Knowledge Interaction Layer Inspired by existing work (Yasunaga et al., 2022; X. Zhang et al., 2022), we use two self-supervised reasoning tasks to pre-train the model, so an interaction layer is designed to achieve the fusion of the two knowledge modalities. The goal is to capture the interdependencies and synergistic relationships between different modalities of knowledge. The interaction layer encodes

the embeddings from question-answer text and knowledge sub-graph separately and combines their representations using special interaction nodes ($\tilde{t}_{int}$, $\tilde{g}_{int}$). It consists of three components: a LTE to encode the aligned question-answer text, $\hat{E}_T$=($\hat{t}_{int}^{(w)}$,$\hat{t}_1^{(w)}$,...,$\hat{t}_I^{(w)}$) (Eq. 10); a GNN to encode the knowledge sub-graph, E_G=($g_{int}^{(l)}$,$g_1^{(l)}$,...,$g_F^{(l)}$) (Eq. 8); and an interaction block to fuse the features of the special interaction nodes. For uniform representation, $\hat{E}_T$ and E_G are re-denoted as $\hat{E}_T$=($\tilde{t}_{int}^{(0)}$,$\tilde{t}_1^{(0)}$,...,$\tilde{t}_I^{(0)}$) and E_G=($\tilde{g}_{int}^{(0)}$,$\tilde{g}_1^{(0)}$,...,$\tilde{g}_F^{(0)}$).

$$(\mathbf{t}_{int},\mathbf{t}_1,...,\mathbf{t}_I),(\mathbf{g}_{int},\mathbf{g}_1,...,\mathbf{g}_F) = \text{Inter}((\tilde{t}_{int}^{(0)},\tilde{t}_1^{(0)},...,\tilde{t}_I^{(0)}),(\tilde{g}_{int}^{(0)},\tilde{g}_1^{(0)},...,\tilde{g}_F^{(0)})) \quad (11)$$

$$(\tilde{t}_{int}^{(j)},\tilde{t}_1^{(j)},...,\tilde{t}_I^{(j)}) = \text{LTE}(\tilde{t}_{int}^{(j-1)},\tilde{t}_1^{(j-1)},...,\tilde{t}_I^{(j-1)}) \quad (12)$$

$$(\tilde{g}_{int}^{(j)},\tilde{g}_1^{(j)},...,\tilde{g}_F^{(j)}) = \text{GNN}(\tilde{g}_{int}^{(j-1)},\tilde{g}_1^{(j-1)},...,\tilde{g}_F^{(j-1)}) \quad (13)$$

$$[\tilde{t}_{int}^{(j)};\tilde{g}_{int}^{(j)}] = \text{Inter-Block}([\tilde{t}_{int}^{(j)};\tilde{g}_{int}^{(j)}]) \quad (14)$$

where $\mathbf{E}_T$=($\mathbf{t}_{int},\mathbf{t}_1,...,\mathbf{t}_I$) and $\mathbf{E}_G$=($\mathbf{g}_{int},\mathbf{g}_1,...,\mathbf{g}_F$) represent the embeddings of question-answer text and knowledge sub-graph after the interaction layer. *Inter-Block* (interaction block) is implemented using a multilayer perceptron, j denotes the index of the interaction layer.

Knowledge Fusion Layer The fusion layer integrates the pre-interaction features of question-answer text and the knowledge sub-graph with post-interaction features, thus preserving individual semantic information while capturing inter-modal interactions. Specifically, for question-answer text, the first token of $\mathbf{E}_T$ after the interaction layer is concatenated with the first token of the aligned $\hat{E}_T$ (Eq. 10):

$$\mathbf{E}_T = (\mathbf{t}_{int}+\hat{t}_{int}^{(w)},\mathbf{t}_1,...,\mathbf{t}_I) \quad (15)$$

For the knowledge sub-graph, the features of all nodes in the $\mathbf{E}_G$ after the interaction layer are concatenated with the corresponding node features of the embedded E_G (Eq. 8):

$$\mathbf{E}_G = (\mathbf{g}_{int},\mathbf{g}_1,...,\mathbf{g}_F)+(g_{int}^{(l)},g_1^{(l)},...,g_F^{(l)}) \quad (16)$$

Finally, to align closely with the methodologies of our base model (Yasunaga et al., 2022), we pretrain our proposed MKE framework by jointly conducting two self-supervised reasoning tasks: masked language modeling and knowledge sub-graph link prediction. The overall training objective for this enhanced approach is defined by the following equation:

$$\mathcal{L}_{Model} = \mathcal{L}_{Mask} + \mathcal{L}_{LinkPred} \quad (17)$$

Experiments

This section analyzes our approach via experiments. We detail the implementation for reproducibility and insight into results. We then compare our model's performance with leading question-answering models and assess individual components through ablation studies. Additionally, we explore the method's effectiveness on various downstream tasks.

Implementation Details

Datasets. We first evaluate our model on two multiple-choice commonsense question-answering datasets: OpenBookQA (OBQA) (Mihaylov, Clark, Khot, & Sabharwal, 2018) and CommonsenseQA (CSQA) (Talmor, Herzig, Lourie, & Berant, 2019). Simultaneously, considering the efficiency during the model training process, we pre-build a multi-modal resource base called **MRB**. MRB is constructed based on the OBQA and CSQA datasets using our proposed knowledge acquisition method, comprising 18,059 pairs of keywords, **18,059 images**, **18,059 description statements**, and **18,059 scene graphs**. Each keyword corresponds to an image, and each image is associated with one description statement and one scene graph. Additionally, we conduct experiments on several other downstream reasoning tasks to demonstrate the model's comprehensive effectiveness, including HellaSwag (Zellers, Holtzman, Bisk, Farhadi, & Choi, 2019), Physical Interaction QA (PIQA) (Bisk, Zellers, Gao, Choi, et al., 2020), and Social Interaction QA (SIQA) (Sap, Rashkin, Chen, Le Bras, & Choi, 2019). For this study, we use the original data split of the above dataset provided in DRAGON (Yasunaga et al., 2022).

Evaluation Metrics. In question-answering tasks, accuracy (Acc.)(%) is a widely used metric for assessing how well a model correctly answers given questions. It is defined as the ratio of the number of questions correctly answered by the model to the total number of questions in the dataset.

Baselines. The models we compare with include DRAGON, GreaseLM (X. Zhang et al., 2022), QA-GNN (Yasunaga, Ren, Bosselut, Liang, & Leskovec, 2021), MHGRN (Feng et al., 2020), and GconAttn (X. Wang et al., 2019), all of which utilize textual knowledge graphs. Additionally, we compare our model with RoBERTa-Large, which does not incorporate any knowledge graph.

Experiment Settings. To ensure rigorous comparison, we adopt the DRAGON model as the base model and perform basic masked language modeling and knowledge sub-graph link prediction pretraining on the same text data with an equal number of training steps. Therefore, the only difference is that our model utilizes multi-modal knowledge and applies corresponding feature processing during the pretraining process. Hyperparameters for pretraining RoBERTa-Base (for image description and scene graph embedding), ViT-Base (for image embedding), and embedding for question-answer text and knowledge sub-graphs are detailed in (Y. Liu et al., 2019), (Dosovitskiy et al., 2021), and (X. Zhang et al., 2022) respectively. Experiments run on an Ubuntu system with 160 GB RAM and four Tesla V100 SXM2 32GB GPUs.

Main Results

We compare the experimental results of our model with various existing state-of-the-art commonsense question-answering models on the development and test datasets of OpenBookQA and CommonsenseQA, as shown in Table 1. We observe consistent improvements in our model compared

Table 1: The table displays the performance comparison of our proposed model with various existing state-of-the-art question-answering models on the development (Dev) and test datasets of OpenBookQA (OBQA) and CommonsenseQA (CSQA).

Model	OBQA Dev Acc. (%)	OBQA Test Acc. (%)	CSQA Dev Acc. (%)	CSQA Test Acc. (%)
RoBERTa-Large	66.8 (±1.1)	64.8 (±2.4)	73.1 (±0.5)	68.7 (±0.6)
GconAttn	64.3 (±1.0)	61.9 (±2.4)	72.6 (±0.4)	68.6 (±1.0)
MHGRN	68.1 (±1.0)	66.9 (±1.2)	74.5 (±0.1)	71.1 (±0.8)
QA-GNN	-	67.8 (±2.7)	76.5 (±0.2)	73.4 (±0.9)
GreaseLM	-	66.9 (±1.0)	78.5 (±0.5)	74.2 (±0.4)
DRAGON	70.8 (±1.3)	72.0 (±0.9)	79.3 (±0.3)	76.0 (±0.5)
Our Model	**75.8 (±0.8)**	**74.9 (±0.6)**	**81.4 (±0.3)**	**78.6 (±0.4)**

to the fine-tuned language model (RoBERTa-Large) and existing language models augmented with textual knowledge graphs (DRAGON, GREASELM, etc.) on both datasets. Specifically, on the OBQA test dataset, our model achieves a 74.9 accuracy, representing a 16% and 4.0% increase over RoBERTa-Large and DRAGON, respectively. Similarly, on the CSQA test dataset, our model achieves a 78.6 accuracy, reflecting a 14% increase compared to RoBERTa-Large and a 3.4% improvement compared to DRAGON. The experimental results indicate that the MKE framework proposed in this paper, which utilizes multi-modal knowledge for enhancement, performs better in supporting auxiliary reasoning.

We also examine our model's performance on complex reasoning problems, as shown in Table 2. Building on previous works (X. Zhang et al., 2022), we categorized complex questions based on the presence of negation (Neg), the presence of conjunction (Coj), the presence of hedge terms (Hed), the number of entity mentions (set at 10, Ent10), and the number of prepositional phrases (set at 3, PP3). Here, the presence of negation or conjunction indicates logical multi-step reasoning, while the presence of hedge terms indicates engagement with complex textual nuance. Additionally, having more entity mentions or prepositional phrases indicates the involvement of more reasoning steps or constraints. From the experimental results, it can be observed that our model outperforms the fine-tuned language model and existing knowledge graph-augmented models significantly across all these categories. Specifically, the negation and conjunction categories show improvements of 3.2% and 4.1%, respectively, over DRAGON, indicating our model's superior logical multi-step reasoning capabilities. The hedge category shows a 5.7% improvement over DRAGON, suggesting that our model can represent more complex textual nuances. The entity mentions and prepositional phrases categories show 6.3% and 5.4% improvements, respectively, over DRAGON, indicating our model handles more reasoning steps or constraints effectively.

Ablation Studies

In this section, we present comprehensive ablation studies to rigorously assess our model's various techniques. We first examine the impact of architectural components on performance. Next, we explore how different modal knowledge affects inference. Finally, we investigate the impact of specific techniques in knowledge fusion on the model's results.

Table 2: Accuracy on OBQA+CSQA dev datasets for questions involving complex reasoning.

Model	Neg	Coj	Hed	Ent10	PP3
RoBERTa-Large	61.7	70.9	68.6	74.5	73.1
QA-GNN	65.1	74.5	74.2	78.6	71.3
GreaseLM	65.1	74.9	76.6	79.4	73.6
DRAGON	75.2	79.6	77.5	83.5	80.9
Our Model	**77.6**	**82.8**	**81.9**	**88.7**	**85.2**

Table 3: Experimental results correspond to the components (MKU, MKF) and their combination (MKU+MKF).

Model	OBQA Acc.	CSQA Acc.
Base	72.0	76.0
Base + MKU	74.1	77.9
Base + MKF	72.6	76.5
Base + MKU + MKF	**74.9**	**78.6**

The Architecture. We conduct comprehensive ablation studies on our framework's components (Table 3). Base represents the base model. MKU represents the utilization of our proposed multiple components for multi-modal knowledge acquisition, embedding, alignment, and interaction. MKF represents the use of our designed component for knowledge fusion without utilizing the multi-modal knowledge proposed in this paper. On the OBQA test dataset, MKU improves the model's accuracy from 72.0 to 74.1, and the MKF component enhances it from 72.0 to 72.6. The combination (MKE) of MKU and MKF further boosts the model's accuracy to 74.9. Similarly, on the CSQA test dataset, MKU boosts accuracy from 76.0 to 77.9, and MKF to 76.5, with MKE enhancing it to 78.6. The combination (MKE) of MKU and MKF further improves the model's accuracy to 78.6. These results validate our MKE framework's effectiveness in improving accuracy.

Multi-modal Knowledge. We assess the effectiveness of different modal knowledge in our multi-modal knowledge framework (see Table 4). Here, E_I denotes image information, E_C denotes image description information, and E_S de-

Table 4: Ablation studies on multi-modal knowledge.

Model	OBQA Acc.	CSQA Acc.
Base	72.0	76.0
Base + E_I	73.2	77.1
Base + E_C	72.8	76.7
Base + E_S	73.0	76.9
Base + E_I + E_C + E_S	**74.1**	**77.9**

notes scene graph information. These experimental results demonstrate the effectiveness of each modality's data. Furthermore, the combined utilization of all modalities yields optimal accuracy performance, as it leverages the fine-grained visual information of the image, such as the spatial relationship between objects, and colors and sizes of objects, as well as the information of the whole image. For specific examples corresponding to Figure 1, qualitative experimental results are depicted in Figure 3. Image information provided global context and detailed size information about characters like Patrick and SpongeBob. Image description information offered fine-grained details like the skin colors of Tom and Jerry. Scene graph information included intricate relational details, such as the connections between Mickey Mouse and Donald Duck with specific items like pants and hats.

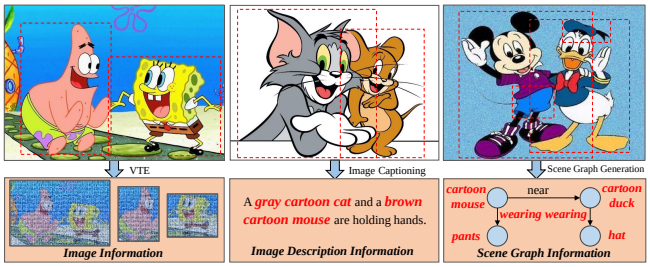

Figure 3: Qualitative analysis of multi-modal knowledge.

Knowledge Fusion. We conducted ablation experiments on the specific techniques of the proposed knowledge fusion component within the MKE framework, as shown in Table 5. Here, TF represents the fusion of question-answer textual information, GF represents the fusion of knowledge sub-graph information, and MKF represents the fusion of both question-answer textual information and knowledge sub-graph information. The experiments are performed under the condition of using existing multi-modal knowledge (Base + MKU). The experimental results demonstrate that on the OBQA test dataset, TF improves the model's accuracy from 74.1 to 74.7, GF enhances the accuracy from 74.1 to 74.5, and MKF improves the model's accuracy from 74.1 to 74.9. On the CSQA test dataset, TF increases the model's accuracy from 77.9 to 78.4, GF enhances the accuracy from 77.9 to 78.2, and MKF raises the model's accuracy from 77.9 to 78.6. These results demonstrate the effectiveness of the techniques within the knowledge fusion component. Both the fusion of question-answer textual and knowledge sub-graph contribute to improving the model's accuracy to a certain extent, and their combination achieves the best experiment results.

Table 5: Ablation studies on knowledge fusion.

Model	OBQA Acc.	CSQA Acc.
Base + MKU	74.1	77.9
Base + MKU + TF	74.7	78.4
Base + MKU + GF	74.5	78.2
Base + MKU + MKF	**74.9**	**78.6**

Table 6: Experiment results on more downstream tasks.

Model	HellaSwag	PIQA	SIQA
RoBERTa-Large	82.3	79.4	75.9
QA-GNN	82.6	79.6	75.7
GreaseLM	82.8	79.6	75.5
DRAGON	85.2	81.1	76.8
Our Model	**87.6**	**83.9**	**78.7**

More Downstream Evaluation Tasks

We also conducted fine-tuning and comprehensive evaluation on several downstream reasoning tasks, with detailed experimental results presented in Table 6. Across various reasoning tasks, including HellaSwag, PIQA, and SIQA, our model consistently outperforms the fine-tuned language model (RoBERTa-Large) and existing knowledge graph-augmented models (QA-GNN, GreaseLM, DRAGON). Notably, for SIQA reasoning tasks, the images are retrieved based on context content, not question-based. Specifically, when compared to DRAGON, our model's accuracy increased by 2.8%, 3.5%, and 2.5% in the HellaSwag, PIQA, and SIQA reasoning tasks, respectively. This underscores our model's robust performance and versatility, highlighting its potential for widespread application in diverse reasoning tasks and laying a strong foundation for its practical utility.

Conclusion

Inspired by the important role of multi-modal knowledge such as vision information in human commonsense question-answering, we propose a commonsense question-answering framework MKE based on the enhancement of multi-modal knowledge, combining textual information with multi-modal knowledge to provide more accurate user-oriented commonsense question answers. We integrate multi-modal knowledge into a CQA system, introduce a parallel embedding technique for embedding this multi-modal knowledge, and employ an alignment-interaction-fusion mechanism to facilitate the seamless integration of this multi-modal knowledge. Extensive experiments on OBQA and CSQA datasets validate our method's effectiveness in both conventional and complex reasoning tasks. Experiments on downstream reasoning tasks further confirm our proposed method's broad effectiveness.

References

Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al. (2020). Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 7432–7439).

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... others (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning* (pp. 2206–2240).

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). A text feature based automatic keyword extraction method for single documents. In *Advances in information retrieval: 40th european conference on ir research, ecir 2018, grenoble, france, march 26-29, 2018, proceedings 40* (pp. 684–691).

Cao, Z., Xu, Q., Yang, Z., He, Y., Cao, X., & Huang, Q. (2022). Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, *35*, 39090–39102.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chen, Q., Xu, G., Yan, M., Zhang, J., Huang, F., Si, L., & Zhang, Y. (2023). Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering. *The 61st Annual Meeting of the Association for Computational Linguistics*.

Chen, Y., Ge, X., Yang, S., Hu, L., Li, J., & Zhang, J. (2023). A survey on multimodal knowledge graphs: Construction, completion and applications. *Mathematics*, *11*(8), 1815.

Cui, W., Lan, Y., Pang, L., Guo, J., & Cheng, X. (2020). Beyond language: Learning commonsense from images for reasoning. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 4379–4389).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Dou, Z.-Y., & Peng, N. (2022). Zero-shot commonsense question answering with cloze translation and consistency optimization. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 10572–10580).

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Feng, Y., Chen, X., Lin, B. Y., Wang, P., Yan, J., & Ren, X. (2020). Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 1295–1309).

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning* (pp. 3929–3938).

He, J., Gutiérrez-Basulto, V., Pan, J. Z., et al. (2023). Buca: A binary classification approach to unsupervised commonsense question answering. In *61st annual meeting of the association for computational linguistics.*

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., & Hajishirzi, H. (2020). Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., & Wen, J.-R. (2022). Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Li, M., Zareian, A., Lin, Y., Pan, X., Whitehead, S., Chen, B., ... others (2020). Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 77–86).

Liu, S., Chang, P., Liang, W., Chakraborty, N., & Driggs-Campbell, K. (2021). Decentralized structural-rnn for robot crowd navigation with deep reinforcement learning. In *2021 ieee international conference on robotics and automation (icra)* (pp. 3517–3524).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

McInnes, B. (2016). Vcu at semeval-2016 task 14: Evaluating definitional-based similarity measure for semantic taxonomy enrichment. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 1351–1355).

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2381–2391).

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nguyen, V.-Q., Suganuma, M., & Okatani, T. (2022). Grit: Faster and better image captioning transformer using dual visual features. In *European conference on computer vision* (pp. 167–184).

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Palta, S., & Rudinger, R. (2023). Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the association for computational linguistics: Acl 2023* (pp. 9952–9962).

Pan, M., Pei, Q., Liu, Y., Li, T., Huang, E. A., Wang, J., & Huang, J. X. (2023). Sprf: A semantic pseudo-relevance feedback enhancement for information retrieval via conceptnet. *Knowledge-Based Systems*, *274*, 110602.

Qin, Y., Cai, Z., Jin, D., Yan, L., Liang, S., Zhu, K., ... others (2023). Webcpm: Interactive web search for chinese long-form question answering.

Rogers, A., Gardner, M., & Augenstein, I. (2023). Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, *55*(10), 1–45.

Rosset, C., Xiong, C., Phan, M., Song, X., Bennett, P., & Tiwary, S. (2020). Knowledge-aware language model pretraining. *arXiv preprint arXiv:2007.00655*.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019). Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4463–4473).

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., ... Zheng, K. (2020). Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 1405–1414).

Sun, Y., Shi, Q., Qi, L., & Zhang, Y. (2022). Jointlk: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 5049–5060).

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of naacl-hlt* (pp. 4149–4158).

Tian, X., Jing, L., He, L., & Liu, F. (2021). Stereorel: Relational triple extraction from a stereoscopic perspective. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 4851–4861).

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. (2017). Graph attention networks. *stat*, *1050*(20), 10–48550.

Wang, M., Wang, H., Qi, G., & Zheng, Q. (2020). Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, *22*, 100159.

Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., ... others (2019). Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 7208–7215).

Xu, G., Chen, H., Li, F.-L., Sun, F., Shi, Y., Zeng, Z., ... Zhang, J. (2021). Alime mkg: A multi-modal knowledge graph for live-streaming e-commerce. In *Proceedings of the 30th acm international conference on information & knowledge management* (pp. 4808–4812).

Xu, Y., Zhu, C., Wang, S., Sun, S., Cheng, H., Liu, X., ... Huang, X. (2021). Human parity on commonsenseqa: Augmenting self-attention with external attention. *arXiv preprint arXiv:2112.03254*.

Yang, H.-Y., & Silberer, C. (2022). Are visual-linguistic models commonsense knowledge bases? In *Proceedings of the 29th international conference on computational linguistics* (pp. 5542–5559).

Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., & Liu, Z. (2022). Panoptic scene graph generation. In *European conference on computer vision* (pp. 178–196).

Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P. S., & Leskovec, J. (2022). Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, *35*, 37309–37323.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *North american chapter of the association for computational linguistics (naacl)*.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4791–4800).

Zhang, M., He, T., & Dong, M. (2024). Meta-path reasoning of knowledge graph for commonsense question answering. *Frontiers of Computer Science*, *18*(1), 181303.

Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Manning, C. D., & Leskovec, J. (2022). Greaselm: Graph reasoning enhanced language models. In *International conference on learning representations.*

Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., ... Yuan, N. J. (2022). Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Zou, A., Zhang, Z., & Zhao, H. (2023). Decker: Double check with heterogeneous knowledge for commonsense fact verification. *The 61st Annual Meeting of the Association for Computational Linguistics*.