2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 6-10, 2024. Sarawak, Malaysia

Enhancing GPT-3.5 for Knowledge-based VQA with In-context Prompt Learning and Image Captioning

Yuling Yang[1,2], Cong Cao[1,2], Fangfang Yuan[1], Shuai Zeng[1,*], Dakui Wang[1,2] and Yanbing Liu[1,2]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{yangyuling, caocong, yuanfangfang, zengshuai, wangdakui, liuyanbing}@iie.ac.cn

***Abstract*—Traditional visual question answering (VQA) often falls short as merely relying on image information is insufficient to answer given questions. Therefore, Knowledge-Based Visual Question Answering (KB-VQA) has emerged. Typically, KB-VQA involves first retrieving knowledge from external knowledge bases, then using the retrieved knowledge in conjunction with the understanding of visual content for joint reasoning to predict answers. However, current models often suffer from weak visual perception capabilities when processing image information. Additionally, due to the incompleteness of external knowledge bases, retrieved knowledge may contain noise or even irrelevant information. Moreover, the re-embedding of knowledge text features during the model's reasoning process may deviate from the original meanings in the knowledge base. To address these challenges, we propose a method for Knowledge-Based Visual Question Answering (KB-VQA) using GPT-3.5, leveraging image captions and in-context prompts. We utilize an advanced captioning model to convert images into accurate textual representations, enhancing the large language model's understanding of visual information. Moreover, we eliminate the need for additional knowledge bases by directly employing GPT-3.5 as a knowledge base for knowledge retrieval and generate logically consistent text during inference to predict answers. Furthermore, we enhance GPT-3.5's question-answering capability for VQA through in-context prompt learning. Experiments on the public OK-VQA dataset demonstrate the superior performance of our model.**

Index Terms—knowledge-based visual question answering, image captioning, in-context learning, large language model

I. Introduction

Visual Question Answering (VQA) is a vision-language task. Given an image and a natural language question about the image, the goal is to provide an accurate natural language answer [1] (as shown in Fig. 1(a)). It combines techniques from both computer vision and natural language processing fields, aiming to enable computer systems to understand and answer questions about the content of images. Thanks to large-scale pre-training on visual-language data, state-of-the-art methods have achieved high performance on several representative benchmarks [2]. Despite the success of these methods, in real-world applications, it is often insufficient to answer the given questions solely based on the information present in the image. Effective utilization of external knowledge resources related to the image is also required. Therefore, Knowledge-based Visual Question Answering (KB-VQA) has gained increasing attention in recent years.

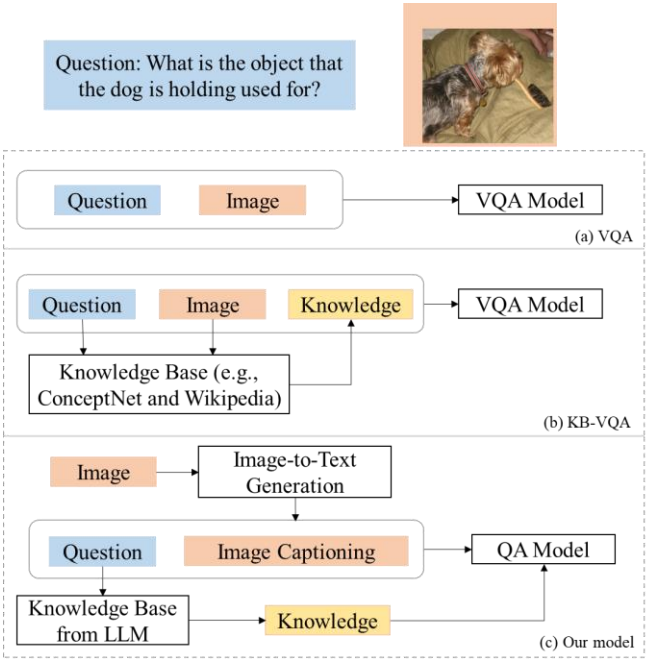

Fig. 1. Comparison between VQA and KB-VQA. (a) Visual Question Answering model, (b) Knowledge-based Visual Question Answering model, (c) our method based on KB-VQA

The conventional approach (as shown in Fig. 1(b)) typically involves first retrieving knowledge from external knowledge bases, such as ConceptNet [3] and Wikipedia [4], and then employing the retrieved knowledge in conjunction with understanding of visual content for joint reasoning to predict answers [5]. However, due to the incompleteness of external knowledge bases, retrieved knowledge may contain noise or irrelevant information. Moreover, re-embedding knowledge text features during the model's reasoning process may deviate from the original meanings in the knowledge base. With the advancements in Large Language Models (LLMs), demonstrating robust task-handling capabilities in many fields, such as information retrieval and open-domain question answering, without relying on external knowledge bases, they can still achieve high performance levels. Inspired by the potential of large language models in knowledge reasoning, we choose to address this challenge by leveraging GPT-3.5, one of the most advanced large language models in natural language processing (NLP). GPT-3.5 encapsulates a vast repository of rich knowledge internally and possesses robust reasoning capabilities. It can provide answers by querying its internal knowledge base and generate logically consistent text during the inference process. Meanwhile, current VQA models often face challenges with weak visual perception

*Corresponding author.

2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-1020-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMC54092.2024.10832004

capabilities when processing image information. To address this challenge, we intend to utilize the advanced image captioning method to convert images into textual representations. This conversion provides easier-to-process inputs, helping models overcome barriers in understanding visual information. This, in turn, enables the model to better comprehend and process the content of image information using natural language processing techniques.

In this paper, we propose a knowledge-based visual question answering model called CPLIC, where we use GPT-3.5 as an implicit knowledge base and unify knowledge retrieval with inference and question-answering steps (as shown in Fig. 1(c)). Specifically, for a given image and question, we first convert the image into a text description format understandable by GPT-3.5 using the captioning model BLIP-2 [6]. Then, we select context examples with similar features to the input through similarity matching. We concatenate the caption-question-answer triplets from these examples to facilitate effective context learning. Finally, we enhance GPT-3.5's ability to predict answers based on questions and image captions using prompt templates.

The contributions of this work are as follows:

- We propose a method for knowledge-based visual question answering using the large language model GPT-3.5, leveraging image captions and in-context prompts for learning.
- We utilize a captioning model to convert images into textual representations and select the most appropriate context examples as prompt information to directly retrieve answers from GPT-3.5 internal knowledge base and generate logically consistent text during inference, which can enhance GPT-3.5's question-answering capability for VQA.
- Extensive experiments on the public OK-VQA dataset demonstrate that our model outperforms baseline methods. This indicates that our model not only effectively retrieves relevant knowledge but also accurately reasons over questions and contexts to predict answers.

II. Related Work

A. Knowledge-based Visual Question Answering

Visual Question Answering (VQA) aims to enable models to comprehend and answer natural language questions about image content, combining techniques from both computer vision and natural language processing. It has been one of the most popular topics in recent years. Compared to VQA, Knowledge-based VQA (KBVQA) places more emphasis on the role of knowledge in the answering process. It leverages structured information from knowledge bases to enhance question understanding, thereby improving answer accuracy and reliability. Zhu et al. [7] proposed using a multimodal heterogeneous graph containing multi-layered information based on visual, semantic, and factual features to describe images and subsequently answer corresponding questions. Marino et al. [8] utilized the powerful implicit reasoning capabilities of Transformer models and integrated symbolized representations from knowledge graphs for answer prediction. Wu et al. [5] introduced a knowledge-based multi-modal answer verification method, utilizing external visual knowledge (via Google image search) and text knowledge in the form of Wikipedia sentences and ConceptNet concepts. The idea is to validate a set of promising answer candidates based on knowledge retrieval specific to each answer. You et al. [9] proposed a Visual Entity Linking (VEL) module, which extracts key entities from images to replace ambiguous references in questions, using explicit entities for entity-oriented queries. They designed a Retriever-Reader framework to retrieve external knowledge for answering purposes.

B. Image Captioning

The image captioning task aims to generate textual descriptions for given images, typically involving the recognition and understanding of content within images, including objects, attributes, and their relationships [10]. Nguyen et al. [11] employed a set of learnable object queries to extract semantics from multi-scale features and input them into a Transformer to generate captions. Wu et al. [12] enhanced visual features by incorporating additional visual information, thereby increasing the contribution of visual cues to caption generation. Zhang et al. [13] not only built a multimodal relationship graph for images but also encoded relationship graphs for all sentences in the dataset to fully capture language features. They then developed a cascaded GAN to achieve cross-domain alignment between images and text pairs, feeding them into a decoder. Zeng et al. [14] aimed to capture hierarchical semantic structures in the text space to facilitate semantic transfer of visual features and generate more fine-grained and coherent phrases and collocations.

C. In-context Learning

With the continuous improvement in the capabilities of large-scale pre-trained language models (LLMs), in-context learning has gradually emerged as a new paradigm in the field of natural language processing. Yang et al. [15] proposed a method using GPT-3 to address the KB-VQA task in a few-shot way, requiring only a small number of VQA examples in context. Monajatipoor et al. [16] attempted to transfer the in-context learning capability from the language domain to the visual-language domain. They initially meta-trained the language model to perform context learning on NLP tasks. Then, they transfer this model to perform the VQA task by attaching a visual encoder. Shao et al. [17] proposed a knowledge-based VQA approach by leveraging answer-heuristic prompts with GPT-3. Two complementary answer heuristics are extracted from the model and encoded as prompts to enhance GPT-3's understanding of the task and thereby augment its capabilities.

III. Methodology

The extensive knowledge stored in large language models (LLMs) and their powerful reasoning abilities enable them to effectively address knowledge-based visual question answering (VQA). Given the abundance of implicit information embedded in the parameters of LLMs, along with their strong reasoning capabilities, our approach uses GPT-3.5 to learn knowledge-based VQA tasks through image captioning and in-context prompts. Unlike fine-tuning pre-trained models to adapt to downstream tasks, our model can adapt to new tasks through in-context prompt learning. It only requires connecting in-context examples to the input

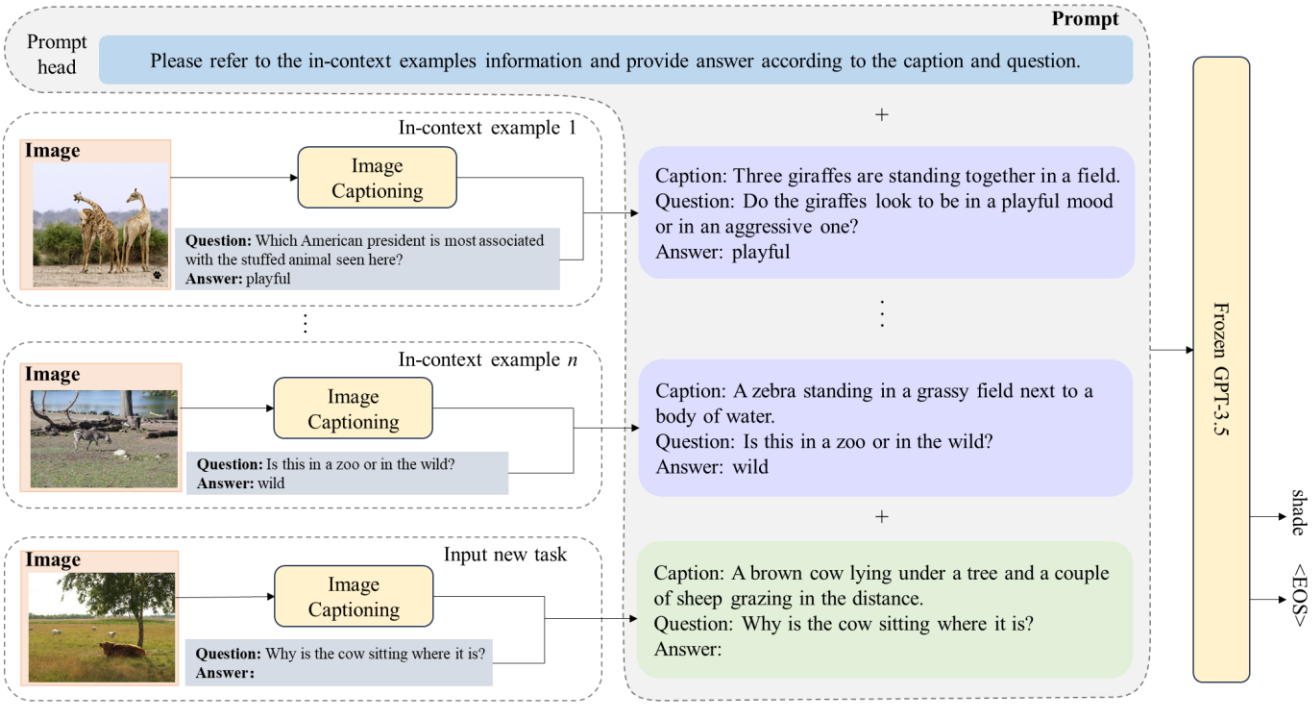

Fig. 2. The overall architecture of our model The blue box represents the prompt head. The purple box is the in-context example (where the image is converted into caption text through an image-to-text generation model). The green box is the input test example (including image caption and question), and the goal is to output predicted answer

during inference as prompts for GPT-3.5, without the need for parameter updates.

A. Problem Formulation

Given a question q and an image v as input, the goal of the knowledge-based VQA is to predict the corresponding answer a by integrating information from a knowledge base. Specifically, since GPT-3.5 cannot directly process images, we utilize in-context learning to formalize novel downstream tasks into text sequence generation tasks during the inference process. Based on the given formatted prompt $p(h, M)$ and the input of the new task x, we predict the target y. In this prompt, h serves as the description of the task prompt header, $M = \{m_1, m_2, \ldots, m_n\}$ represents n in-context examples in the new task, where $m_i = (x_i, y_i)$ represents the input-target pair of the task. The input x comprises the image-to-text caption c and the question q. The target output y is represented as a text sequence consisting of T tokens, denoted as $y = (y^1, y^2, \ldots, y^T)$. Therefore, at each decoding step t:

$$\hat{y}^t = \underset{y^t}{\operatorname{argmax}}\, p_{LLM}(y^t \mid p, x, \hat{y}^{<t}) \tag{1}$$

B. Overall Framework

Fig. 2 illustrates the overall framework of our model CPLIC, which prompts GPT-3.5 to complete the VQA task using a constructed prompt template. Specifically, the prompt template consists of a task description prompt header h, a set of caption-question-answer triplets, the input image caption c, and the question q. The prompt header h (as shown in the blue box in Fig. 2) is a fixed string and marks the beginning of the entire prompt. The purple box in Fig. 2 represents n in-context examples $\{x_i, y_i\}_{i=1}^{n}$. Then, we concatenate the caption-question-answer triplet M and the input x of the new task to construct the complete prompt. For a new task, as shown in the green box in Fig. 2, we first translate the image into text form using a captioning model, converting the image-question pair into text-question format. The input x is the concatenated string of the image caption and the question. GPT-3.5 takes the constructed prompt text as input and implicitly retrieves and infers knowledge from the language model, with the target y being the output corresponding to the answer.

C. Image Captioning

Since GPT-3.5 inherently lacks understanding of image information, to provide the model with sufficient visual context, we use a caption generation model to convert images into text formats comprehensible by GPT-3.5. BLIP-2 [6] is a multimodal pre-training model trained under the condition of freezing both the image encoder and the text encoder. Its training consists of two stages. The first stage involves joint visual encoding, freezing the parameters of the Image Encoder for training to obtain high-quality alignment vector representations of image-text pairs. In the second stage, the visual encoder is combined with a large language model, and the parameters of the LLM are frozen for training to enable the model to acquire strong zero-shot capabilities and the ability to generate text from images. Compared to previous image-to-text generation models, BLIP-2 adopts an instruction-recognition approach to extract visual features, which adequately captures image information, resulting in more comprehensive captions describing the images. Specifically, for each image-question pair, we first use the

image-language model BLIP-2 to convert every image into a caption form, and then use the question information caption as context information to achieve effective in-context learning. In this way, GPT-3.5 can give the predicted answer according to the question and text description through context prompt learning.

D. In-context Example

It turns out that providing more in-context examples to the model leads to better performance [15]. To better utilize these context examples, we adopt the following two approaches to obtain examples.

In-context Example Selection. In-context example selection aims to search for the optimal examples among all available instances at each inference time input. In simple terms, we choose in-context examples that share similar question features with the input x. Specifically, given an inference-time question, we utilize the text encoder of the CLIP model [18] to extract its textual features and calculate its cosine similarity with the questions in the in-context examples. Subsequently, we average the textual similarity of the questions with the visual similarity of the images to guide the selection of examples. We then select the top-k questions with the highest similarity and use the corresponding examples as in-context instances. This approach ensures that the selected examples closely align with the characteristics of the input, enhancing the relevance and effectiveness of the in-context prompts.

Prompt Integration. Given an input x, we use $n * k$ in-context examples to generate k prompts. Through this approach, we prompt GPT-3.5 k times to obtain k answers, where k is the number of queries to integrate. Among the k answer predictions, we choose the one that appears most frequently as the final answer. This strategy helps reduce the uncertainty caused by randomness, enhancing the stability and reliability of the model outputs. Additionally, by integrating multiple integrations, we can comprehensively consider the model predictions under different prompts, thus better reflecting the model's performance in various in-context scenarios.

E. Retrieval and Inference

We concatenate the triplet M of captions, questions, and answers with the new task input x to form a complete prompt input for GPT-3.5. The construction of the prompt template aims to provide the model with sufficient information to effectively utilize in-context prompt learning and enhance the model's understanding ability. In this process, we do not need to rely on additional external knowledge bases but directly retrieve relevant knowledge information from within the model. This internal retrieval approach not only simplifies the system's complexity but also improves the logical consistency during the reasoning phase. Additionally, leveraging the powerful reasoning capability of large language models allows for efficient prediction of answers. The goal is to output answers corresponding to the given image and question, thereby achieving knowledge-based visual question answering tasks.

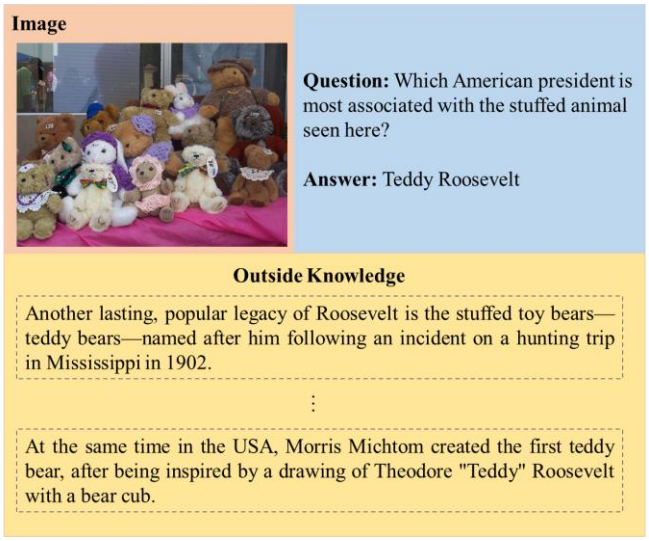

Fig. 3. An example of OK-VQA dataset, where the questions require external knowledge resources to be answered.

IV. Experiment

A. Dataset

We evaluate our model on the OK-VQA dataset [19], which is currently the largest knowledge-based VQA dataset. The OK-VQA dataset consists of 14,055 questions related to 14,031 images from the COCO dataset [20]. The questions in the dataset cover various knowledge categories, and each data sample consists of a question, a corresponding image, and 10 factual answers. This dataset is designed to evaluate the ability of models to answer questions about an image, where questions are manually filtered to ensure that external knowledge is required to answer relevant questions. As shown in Fig. 3, this is an example from the OK-VQA dataset, which asks about the relationship between the stuffed animal and the U.S. president. It is obvious that the visual information of the image is not sufficient to answer this question. Not only does the model need to identify the specific type of stuffed animal, but it also needs to connect the image content with external knowledge resources to retrieve information about teddy bears and U.S. presidents to get the answer.

B. Overall Performance

We select several classic models as baselines for comparison with CPLIC, including Q-only [19], BAN + KG + AUG [21], Mucko [7], ConceptBert [22], KRISP [8], MAVEx[5], UnifER[23], PICa[15], HPM [24]. Most of these methods utilize external knowledge resources (e.g., ConceptNet [3] and Wikipedia [4]) for explicit knowledge retrieval, while a few methods employ the pre-trained language models to acquire and exploit knowledge. We did not introduce additional external resource libraries and solely used GPT-3.5 as the knowledge resource repository. Table I presents the performance results of various methods on the OK-VQA dataset. It can be observed that our results outperform all compared methods. This demonstrates the superior performance of GPT-3.5 in implicit knowledge retrieval and reasoning compared to explicit retrieval of external knowledge. Compared to GPT-3, which also uses internal knowledge base retrieval, our method has a certain improvement.

TABLE I. COMPARISON RESULTS OF CPLIC WITH BASELINE METHODS

Model	Knowledge Resources	Accuracy(%)
Q-only	-	14.93
BAN + KG + AUG	Wikipedia + ConceptNet	26.70
Mucko	Dense Captions	29.20
ConceptBert	ConceptNet	33.66
KRISP	Wikipedia + ConceptNet	38.90
MAVEx	Wikipedia+ConceptNet+Google Images	40.28
UnifER	ConceptNet	42.13
PICa	GPT-3	48.00
HPM	ConceptNet	49.36
CPLIC(ours)	GPT-3.5	**51.68**

C. *Ablation Study*

To verify the effectiveness of key components of CPLIC, we conduct ablation experiments on the OK-VQA dataset, aiming to analyze various essential components to assess their impact on model performance.

Image-to-text Generation. The task requires the model to generate textual descriptions of the visual content in images. To verify the performance of our selected image captioning model, we chose two models, VinVL [25] and BLIP [26] for comparison. We fine-tune the models for the image captioning task, and the results are shown in Table II. It shows that BLIP-2 performs better, indicating that the captions provide a more accurate description of the image content.

In-context example selection. Considering that different methods of selecting in-context examples may affect the in-context information in the final prompt template, we chose three methods to verify the impact of in-context example selection on the performance of our model. These methods are "Random", "Question" and "Question & Image".

- **Random**. In-context examples are randomly selected.
- **Question.** In-context examples are selected based on the textual similarity of questions.
- **Question & Image.** In-context examples are selected based on the cosine similarity between the textual features obtained from the text encoder of the CLIP model [18] and all available in-context examples' questions. Then, the question text similarity is averaged with the image visual similarity to guide the selection of examples.

Number of In-context examples. To validate the impact of the number of in-context examples n on model accuracy, we fix the method of selecting in-context examples and vary the number of examples within the same selection method to evaluate performance of the model.

TABLE II. ABLATION ON IMAGE CAPTIONING METHODS.

Image Captions	VinVL	BLIP	BLIP-2
Accuracy (%)	47.67	49.50	**50.62**

TABLE III. COMPARISON OF THREE CONTEXT SELECTION METHODS AND RESULTS OF THE NUMBER OF IN-CONTEXT EXAMPLES

Methods	Feature	n				
		1	8	16	24	32
Random	Random selection	37.09	43.98	46.76	47.77	48.37
Question	Average question similarity	37.75	45.56	47.91	49.27	49.97
Question & Image	Average question and image similarity	38.24	46.82	48.87	49.96	50.62

Table III shows the results of context example selection using different selection methods and varying numbers of context examples. The experiments indicate that among the three selection methods, randomly selecting context examples does not benefit accuracy improvement. Conversely, considering both question text similarity and image visual similarity to retrieve context examples can enhance accuracy. It indicates that optimizing model performance can be achieved by balancing considerations of both textual and visual features when selecting in-context examples. Under fixed selection methods, the model's performance improves with an increase in the number n of in-context examples, and the growth rate stabilizes gradually. As shown in Fig. 4, the growth rate of the entire curve significantly increases with the increase in the number of context examples, and then gradually stabilizes.

Effect of Multi-prompt Integration. Multi-prompt integration allows the model to utilize more in-context examples to enhance its performance during inference. To assess the impact of prompt templates on model accuracy, we set the selection method for in-context examples as "Question & Image", which means selecting in-context examples based on the similarity between questions and images. We set the number of in-context examples n to 32. As shown in Fig. 5, the change in accuracy with the increase in the number of prompts k during the in-context learning process. Experimental results demonstrate that using multiple prompts leads to an improvement in model accuracy. This integration strategy effectively utilizes a greater number of in-context examples, providing the model with richer information, thus enhancing both performance and generalization capability.

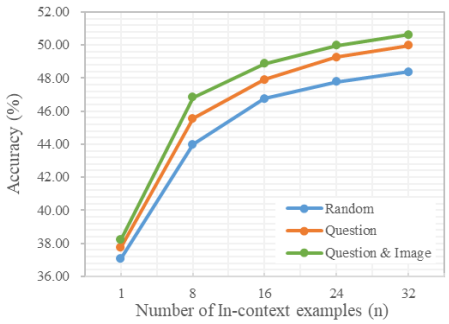

Fig. 4. Result on the number of in-context examples

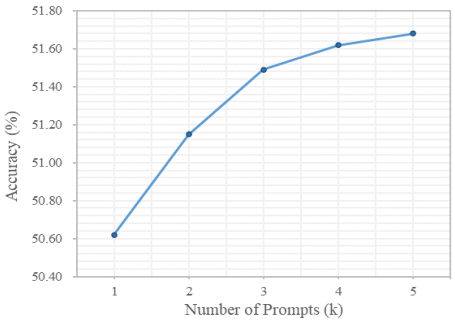

Fig. 5. Multi-prompt Integration performance on OK-VQA. Comparison of the accuracy and number of prompts used during in-context learning

V. Conclusion

This paper introduces a knowledge-based visual question answering model called CPLIC, which utilizes the large language model GPT-3.5 and learns through image captioning and in-context prompts. By employing a captioning model to convert images into textual form, CPLIC enhances the understanding of visual information. CPLIC can directly utilize GPT-3.5 as a knowledge base for knowledge retrieval to predict answers. Additionally, we enhance the question-answering capabilities of GPT-3.5 for VQA through in-context prompt learning. Our next step involves further exploring the integration of explicit and implicit knowledge, adopting more advanced methods for comprehensive image information extraction, and leveraging other large language models.

Acknowledgment

This research is supported by the National Key R&D Program of China (No. 2023YFC3303800).

References

[1] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[J]. Advances in neural information processing systems, 2022, 35: 23716-23736.

[2] Bao H, Wang W, Dong L, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts[J]. Advances in Neural Information Processing Systems, 2022, 35: 32897-32912.

[3] Liu H, Singh P. ConceptNet—a practical commonsense reasoning tool-kit[J]. BT technology journal, 2004, 22(4): 211-226.

[4] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[5] Wu J, Lu J, Sabharwal A, et al. Multi-modal answer validation for knowledge-based vqa[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(3): 2712-2721.

[6] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.

[7] Zhu Z, Yu J, Wang Y, et al. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 1097-1103.

[8] Marino K, Chen X, Parikh D, et al. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14111-14121.

[9] You J, Yang Z, Li Q, et al. A retriever-reader framework with visual entity linking for knowledge-based visual question answering[C]//2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023: 13-18.

[10] Song P, Guo D, Zhou J, et al. Memorial GAN With Joint Semantic Optimization for Unpaired Image Captioning[J]. IEEE transactions on cybernetics, 2023, 53(7): 4388-4399.

[11] Nguyen V Q, Suganuma M, Okatani T. Grit: Faster and better image captioning transformer using dual visual features[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 167-184.

[12] Wu M, Zhang X, Sun X, et al. Difnet: Boosting visual information flow for image captioning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18020-18029.

[13] Zhang W, Shi H, Guo J, et al. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 3335-3343.

[14] Zeng P, Zhu J, Song J, et al. Progressive tree-structured prototype network for end-to-end image captioning[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 5210-5218.

[15] Yang Z, Gan Z, Wang J, et al. An empirical study of gpt-3 for few-shot knowledge-based vqa[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 3081-3089.

[16] Monajatipoor M, Li L H, Rouhsedaghat M, et al. MetaVL: Transferring In-Context Learning Ability From Language Models to Vision-Language Models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023: 495-508.

[17] Shao Z, Yu Z, Wang M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14974-14983.

[18] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[19] Marino K, Rastegari M, Farhadi A, et al. Ok-vqa: A visual question answering benchmark requiring external knowledge[C]//Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2019: 3195-3204.

[20] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[21] Li G, Wang X, Zhu W. Boosting visual question answering with context-aware knowledge aggregation[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1227-1235.

[22] Gardères F, Ziaeefard M, Abeloos B, et al. Conceptbert: Concept-aware representation for visual question answering[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 489-498.

[23] Guo Y, Nie L, Wong Y, et al. A unified end-to-end retriever-reader framework for knowledge-based vqa[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 2061-2069.

[24] Sun Z, Hu Y, Gao Q, et al. Breaking the Barrier Between Pre-training and Fine-tuning: A Hybrid Prompting Model for Knowledge-Based VQA[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 4065-4073.

[25] Zhang P, Li X, Hu X, et al. Vinvl: Revisiting visual representations in vision-language models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 5579-5588.

[26] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International conference on machine learning. PMLR, 2022: 12888-12900.