Solution for SMART-101 Challenge of ICCV Multi-modal Algorithmic Reasoning Task 2023

Xiangyu Wu¹², Yang Yang¹, Shengdong Xu¹, Yifeng Wu², Qingguo Chen², Jianfeng Lu¹,

¹Nanjing University of Science and Technology

Abstract

In this paper, we present our solution to a Multimodal Algorithmic Reasoning Task: SMART-101 Challenge. Different from the traditional visual question answering datasets, this challenge evaluates the abstraction, deduction, and generalization abilities of neural networks in solving visuolinguistic puzzles designed specifically for children in the 6-8 age group. We employed a divide-andconquer approach. At the data level, inspired by the challenge paper, we categorized the whole questions into eight types and utilized the llama-2-chat model to directly generate the type for each question in a zero-shot manner. Additionally, we trained a yolov7 model on the icon45 dataset for object detection and combined it with the OCR method to recognize and locate objects and text within the images. At the model level, we utilized the BLIP-2 model and added eight adapters to the image encoder VIT-G to adaptively extract visual features for different question types. We fed the pre-constructed question templates as input and generated answers using the flan-t5-xxl decoder. Under the puzzle splits configuration, we achieved an accuracy score of 26.5 on the validation set and 24.30 on the private test set.

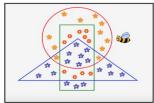
1. Introduction

Visual question answering (VQA) [10, 9] task requires the model to take both the question Q in natural language and the image I as input and generate the answer A according to the information contained in the inputs. With the development of multi-modal large language model [8] technology, these models have demonstrated significant effectiveness in answering questions that require complex logical reasoning abilities.

Traditional visual question answering tasks primarily focus on real-world scene datasets, which evaluate the deep model's ability to recognize and locate objects in images and questions. By performing simple feature fusion, the model can provide accurate answers. As shown in Figure 1, the difficulty in this challenge is to evaluate the generalization abilities of deep neural networks in solving visuolinguistic puzzles designed specifically for children in the 6-8 age group and to understand the algorithmic reasoning abilities of SOTA deep models. This SMART-101 [3] dataset consists of 101 unique puzzles that require a mix of several elementary skills, including arithmetic, algebra, and spatial reasoning, among others. The currently deep models offer reasonable performances on puzzles in a supervised setting, they are not better than random accuracy when analyzed for generalization, and fail entirely on out-of-distribution generalization when the training and testing sets are disjoint at the puzzle levels.



Q: What is in the water? A: boat.



Q: A bee collected pollen from all the flowers inside the rectangle but outside the triangle. From how many flowers did the bee collect pollen? Options: A: 9, B: 10, C: 13, D: 17, E: 20 A: A

Figure 1: The difference between traditional VQA and SMART-101 challenge.

To address these challenges, we propose a divide-and-conquer approach. Firstly, we noticed the remarkable capability of large language models in zero-shot settings. Therefore, inspired by the SMART-101 challenge paper, we categorized the questions into eight types and used the llama-2-chat [11] model to directly generate the question type by constructing a proper prompt template in a zero-shot manner. Secondly, we observed that the object icons in the

²Alibaba International Digital Commerce Group

competition dataset are sourced from the icon45 dataset. Therefore, we automatically constructed an object detection dataset by randomly adding icons on a whiteboard background. Then, we trained a yolov7 [12] object detector and combined it with OCR [7] methods to locate and recognize objects and text in the images. Lastly, we employed the BLIP-2-flan-t5-xxl [8] model, which is a multi-modal large language model with strong capabilities in visual understanding and text generation. We added eight adapters to the image encoder VIT-G [5] to adaptively extract visual features for different question types. We combined this information with candidate options to construct a question template, which was then passed as input to the BLIP-2 model to generate the final answer.

We introduce a divide-and-conquer approach, and its contributions can be summarized as follows:

- We propose the divide-and-conquer approach, utilizing large language models in a zero-shot paradigm to directly predict the type of each question.
- We automatically created an object detection dataset and trained an object detector and OCR model to locate and recognize objects and text in the images.
- We adopted the BLIP-2 multi-modal large language model and added multiple adapters to extract different visual features. Leveraging the powerful visual understanding and text generation capabilities of BLIP-2 to predict the answers.

2. Related Work

2.1. Large Language Model

Large Language Models (LLMs) are AI models that can understand and generate human-like text. They are trained on a vast amount of text data and have been used in a variety of applications, such as translation, summarization, and coding. Some notable examples of LLMs include Chatgpt [6] by OpenAI and flan-t5 [4] by Google. However, these models also pose challenges, including potential misuse and the difficulty of controlling their output. Despite these challenges, the field of LLMs is advancing rapidly, with ongoing research aimed at improving their capabilities and addressing their limitations.

2.2. Vision Adapter

Vision Adapters [1, 2] are a recent development in the field of AI that aims to enhance the capabilities of existing models. For instance, the Vision Transformer Adapter is the first multi-task vision transformer adapter that learns generalizable task affinities which can be applied to novel tasks and domains. Another research is the CLIP-Adapter [2] proposed by Peng Gao et al., which conducts fine-tuning with feature adapters on either visual or language branches. This approach has shown significant progress in visual rep-

resentation learning. These adapters are integrated into an off-the-shelf vision transformer backbone and can simultaneously solve multiple dense vision tasks in a parameter-efficient manner. They outperform not only the existing convolutional neural network-based multitasking methods but also the vision transformer-based ones.

2.3. Multi-modal Large Language Model

In light of this complementarity, unimodal LLMs and vision models run towards each other at the same time, ultimately leading to the new field of Multi-modal Large Language Model (MLLM) [8, 13, ?]. Formally, it refers to the LLM-based model with the ability to receive and reason with multi-modal information. It is able to receive and reason with information from multiple modalities, including text, images, and speech. MLLMs offer several advantages over unimodal LLMs, including more human-like perception, a more user-friendly interface, and the ability to support a larger range of tasks. MLLMs are seen as a potential step forward in the development of Artificial General Intelligence.

3. Method

The overall architecture of our approach is illustrated in Figure 2, consisting of four components: original input, information extraction, model input, and model with adapter.

3.1. Original input

The dataset for this challenge consists of an abstract image and a text input. The image is not a representation of a natural scene but rather an abstract composition of elements such as numbers, lines, boxes, and icons. The text input consists of a question and five options, with the questions being relatively complex and requiring strong logical reasoning abilities to generate correct answers.

3.2. Information extraction

Providing sufficient visual and textual information to the multi-modal model is crucial when dealing with complex multiple-choice visual question answering tasks. The dataset for this competition consists of 101 questions, categorized into eight types: counting, arithmetic, algebra, spatial reasoning, measuring, logic, and pathfinding. Each question type requires different visual information and reasoning abilities to answer. To better address these challenges, we employ a large language model using the zeroshot paradigm to predict the type of each question and reclassify the 101 questions accordingly. Specifically, we construct a prompt template as the input for the llama-2-chat [11] model, which states: "There are eight question types: [counting, arithmetic, algebra, spatial reasoning, measuring, logic, path finding]. So, question which

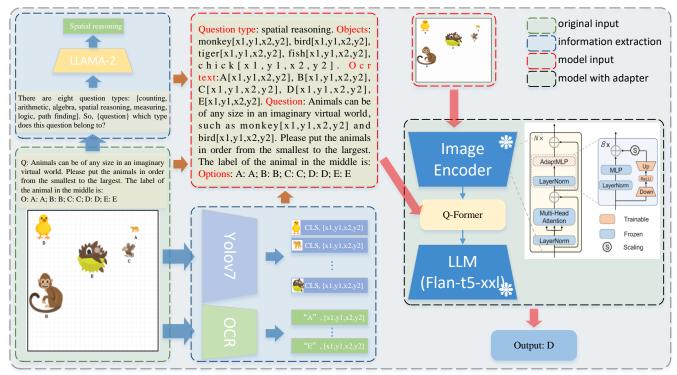


Figure 2: In the framework. "Original input" represents the raw image and question. "Information extraction" represents performing object detection, OCR, and generating the question type on the raw image and question. "Model input" consists of a question template and an image. "Model with adapter" refers to the addition of trainable adapters on the image encoder VIT-G of BLIP-2.

type does this question belong to?". By utilizing this input format, the model can classify the questions based on their content and context. Considering the uncertainty in the large language model's outputs. For each question, we randomly selected 100 samples from all 2000 samples for a certain question for prediction. By aggregating the most frequently occurring output from multiple predictions, we determine the type of question. Through this method, we gain a better understanding of question types and provide more precise guidance for answering questions. This is crucial for successfully completing multiple-choice visual question answering tasks as each question type requires different reasoning and problem-solving approaches.

For this competition dataset, we noticed that it contains various elements in the images, such as numbers, English letters, animals, plants, and everyday objects. The presence of these elements inspired us to utilize object detection and OCR (Optical Character Recognition) [7] techniques for locating and recognizing text and icons in the images. To achieve this goal, we trained an object detector using the icon45 dataset. Specifically, we randomly selected n icons, each with a randomly assigned size, and placed them randomly on a whiteboard. This process allowed us to create a simple dataset for icon object detection. Next, we loaded the pre-trained weights of YOLOv7 [12] and trained an ob-

ject detector capable of recognizing and localizing various icons in the images. This object detector helps us find and identify different icons in the images. For extracting textual information from the images, we employed the PaddleOCR [7] model for text recognition. This model assists in recognizing the textual content within the images and determining its position in the image. By combining object detection and OCR techniques, we can effectively extract important information from the images and utilize it for next processing and analysis.

3.3. Model input

Through the second step, we obtained the type of each question, as well as the category, position of the icons, and text content on the image. Then, we constructed a text template and input it into the BLIP-2 [8] model. This template includes the type of question, all objects on the image (including their categories and positions), and the text content and position recognized by OCR. For example, as shown in Figure 2, for a spatial reasoning type question, we construct the following template: "Question type: spatial reasoning. Objects: monkey[x1,y1,x2,y2], bird[x1,y1,x2,y2], tiger[x1,y1,x2,y2], fish[x1,y1,x2,y2], chick[x1,y1,x2,y2], Ocr text:A[x1,y1,x2,y2], B[x1,y1,x2,y2], C[x1,y1,x2,y2], D[x1,y1,x2,y2], E[x1,y1,x2,y2]. Question: Animals can

be of any size in an imaginary virtual world, such as monkey[x1,y1,x2,y2] and bird[x1,y1,x2,y2]. Please put the animals in order from the smallest to the largest. The label of the animal in the middle is: Options: A: A; B: B; C: C; D: D; E: E." This template includes the type of question (spatial reasoning), all objects on the image (monkey, bird, tiger, fish, and chick) and their positions, text content (A, B, C, D, and E) recognized by OCR and their positions, as well as the specific description and options of the question. Then, we input this template together with the corresponding image into the BLIP-2 model for answer prediction. This method combines rich visual information and text information to effectively handle complex multiple-choice visual question answering tasks.

3.4. Model with adapter

In this competition task, we found that different types of questions require the extraction of different visual features for prediction. However, if a separate model is trained for each type of question, the cost would be very high. Therefore, we adopted a more efficient method. Specifically, we adopted the idea of adapters and added 8 visual adapters [1] to the image encoder VIT-G [5] of the blip-2 [8] model. Each adapter corresponds to a certain type of question and can adaptively extract the visual information needed for that type of question. In this way, we can handle various types of questions under a unified framework without having to train a separate model for each type of question. This method can not only greatly reduce the training cost but also improve the generalization ability of the model. Because each adapter is trained on the same model, they can share some parameters and structures of the model, thereby improving the generalization ability of the model.

4. Experiment

Dataset. The competition dataset is provided by the official organizers, which includes both a training set and a test set. The training data consists of 101 unique puzzles, with each puzzle having 2000 instances. Each puzzle sample contains an image, a question, and corresponding options. For each type of split, we will evaluate submissions on 100 puzzles. Furthermore, the new puzzle instances may originate from root puzzles that have never been encountered before.

Metric. In order to evaluate the model's exceptional generalization capability on SMART, we employ the segmentation method outlined in the original paper, referred to as Puzzle Split. This entails assessing novel, previously unobserved root puzzle instances that demand the same foundational skills as those in the training set. Recognizing the diverse difficulty levels of the puzzles, we also incorporate weights for each puzzle. Our chosen evaluation metric is

Method	acc	text_wosa	vl_wosa	tot_wosa
random	0.93	0.00	2.72	1.59
baseline	18.69	16.35	21.09	19.12
our	24.30	24.04	21.77	22.71

Table 1: Private test set result.

the Weighted Option Selection Accuracy (WOSA), which is quantified by the following formula:

$$100 \times \frac{\sum_{i=1}^{N} w_i acc_i}{\sum_{i=1}^{N} w_i}$$
 (1)

where w_i represents the weight of each puzzle in the test set, where acc_i is 1 if the answer is correct, and 0 otherwise.

Implementation Detail. In our study, we trained our method based on BLIP-2 model. The training was conducted using 8 A100 GPUs, and the optimal performance was achieved at epoch 5. The learning rate was set to 3e-5, with a batch size of 32 per GPU.

Result on private test set. Table 1 shows the performance of our method on the private test set. From the results, compared to the baseline model in the original paper, our method has improved by 5.61% on the accuracy (acc) metric. Due to the powerful text reasoning ability of the large language model, it has increased by 7.69% on the text_wosa metric. In addition, there are also significant improvements in the vl_wosa and total tot_wosa metrics.

Ablation Study on evaluation set. To analyze the contribution of each component of our method, we conduct more ablation studies on the evaluation set of the competition. From Table 2, compared to the blip-2-flan-t5-xxl model, the visual adapter has made a significant improvement. This is due to the extraction of different visual information for different types of questions. In addition, the text information on the image and the category and position of icons of the image also provides more abundant visual information for the model.

Method	ACC
BLIP-2-Flan-t5-xxl	21.1
+ adapter	24.3
+ OCR	25.1
+ YOLOv7	26.5

Table 2: Ablation experiment.

5. Conclusion

In this competition, we found that providing sufficient input information to the model is crucial. This includes the type of question, as well as the category, position, and text content of the icons on the image. We used a large language model to predict the type of each question through a zeroshot paradigm and constructed a text template to input into the model. In addition, we also used object detection and OCR technology to locate and recognize the text and icons on the image. Finally, we input all this information into the model for answer prediction. This method, which combines rich visual information and text information, can effectively handle complex multiple-choice visual question answering tasks and has achieved good results.

References

- Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. May 2022.
- [2] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. May 2022.
- [3] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin Smith, and JoshuaB. Tenenbaum. Are deep neural networks smarter than second graders? Dec 2022.
- [4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models, 2022. https://arxiv.org/pdf/2210.11416.pdf.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, Oct 2020. https://arxiv.org/pdf/2010.11929.pdf.
- [6] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling, 2021. https://arxiv.org/pdf/2103.10360.pdf.
- [7] Baidu Inc. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system, 2022. https://arxiv.org/pdf/2206.03001.
- [8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. https://arxiv.org/pdf/2301.12597.pdf.
- [9] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. ICLR, Jan 2022.
- [10] Alireza Salemi, Mahta Rafiee, and Hamed Zamani. Pretraining multi-modal dense retrievers for outside-knowledge visual question answering. Jun 2023.
- [11] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, CristianCanton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes,

- Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael, Smith Ranjan, Subramanian Xiaoqing, Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, JianXiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stoinic, Sergey Edunov, Thomas Scialom, and Meta Genai. Llama 2: Open foundation and fine-tuned chat models, 2023. https://arxiv.org/pdf/2307.09288.pdf.
- [12] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2023. https://arxiv.org/pdf/2207.02696.pdf.
- [13] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022. https://arxiv.org/pdf/2202.03052.