Look, Remember and Reason: Visual Reasoning with Grounded Rationales

Apratim Bhattacharyya [1] Sunny Panchal [1] Mingu Lee [1] Reza Pourreza [1]
Pulkit Madan [1] Roland Memisevic [1]
[1]Qualcomm AI Research, an initiative of Qualcomm Technologies Inc.

Abstract

The ability of Large language models (LLM) models to perform complex visual reasoning has not yet been studied in detail. Here, we address the key challenge that in many visual reasoning tasks, the visual information needs to be tightly integrated in the reasoning process. We draw inspiration from human visual problem solving which can often be cast as the three step-process of "Look, Remember, Reason": where visual information is incrementally extracted using low-level visual routines in a step-by-step fashion. To this end, we introduce rationales that allow us to integrate low-level visual capabilities, such as object recognition and tracking, as surrogate tasks. We show competitive performance on diverse visual reasoning tasks from the CLEVR, CATER, and ACRE datasets over state-of-the-art models designed specifically for these tasks.

1. Introduction

Autoregressive large language models (LLMs) have shown impressive results on various reasoning tasks such as on grade school math problems [2] and even on LSAT [10]. Language models designed for these problems process only textual data to reason and solve the target task. Many real-world scenarios, however, require humans to reason in complex domains that engage various heterogeneous sensory inputs, *e.g.*, perceptual cues and language. Motivated by this, multimodal LLMs [1, 9, 18] have gained traction, which model information both from the textual and the visual domains. While these models perform well on the tasks that rely on the global visual-textual relationships, *e.g.*, captioning or dialogue [1, 4, 9], the ability of multimodal LLMs to understand spatio-temporal relationships and causal structures in visual data is rather under-explored.

Consider the visual reasoning problem such as in Fig. 1 from ACRE [14], where the objective is to correctly answer whether the query objects (in bottom left) would activate the "Blicket" machine. Humans can solve this problem through a multi-step reasoning process where we attend to and extract

Figure 1. Our "Look, Remember, Reason" (LRR) model solves complex visual reasoning problems by generating grounded rationales with surrogate tasks, *e.g.*, object re-identification, to enable necessary low-level visual capabilities. Our model "looks" at the visual input to extract relevant low-level information step-by-step, and it "remembers" results of intermediate steps. In the above example, this allows our LRR model to "reason" whether the query objects could activate the "Blicket" machine.

visual information step by step using our low-level visual capabilities, such as object recognition and re-identification. For example, one strategy that humans may follow to solve this problem is: read the question; inspect the scene to create an overview of the present objects as well as any relevant low-level visual information; memorize the relevant information along the way; finally state the answer based on the extracted information. Such a reasoning process is crucial to deal with both the complexity of the task and the need to filter the rich visual data for relevant information. In short, such a reasoning process can be thought of as consisting of the three intermediate sub-tasks "Look, Remember, Reason" – looking for relevant visual cues, remembering the relevant

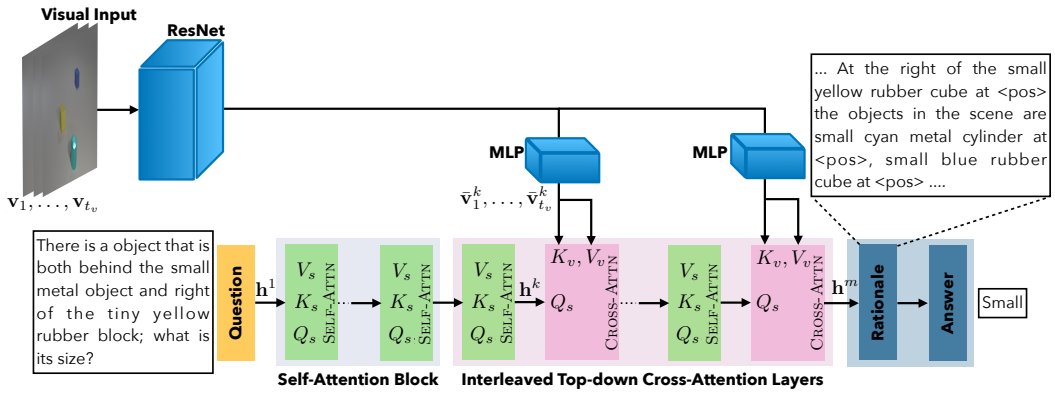

Figure 2. The architecture of our LRR model, highlighting the use of interleaved top-down cross-attention layers in between self-attention layers higher up in the hierarchy.

cues along the way, and finally aggregating the collected information to arrive at the final answer. In this work, we boost the uni-modal large language models for texts to perform general-purpose multimodal visual reasoning by augmenting them with *low-level visual capabilities*.

Our key contributions are: 1. We equip an off-the-shelf language model with the low-level visual capabilities to solve a diverse range of visual reasoning tasks. This is accomplished by training the LLM indirectly using surrogate tasks expressed in natural language requiring the generation of relevant rationales that follow the paradigm of "Look, Remember, Reason" and are grounded in the visual input. 2. We show that it is crucial in these tasks to let high-level concepts modulate the perceptual pathway, and we present an adapter module that accomplishes this through top-down attention controlled by the LLM. 3. Our general-purpose LRR model can perform varied visual reasoning tasks, including spatial reasoning (CLEVR;[7]), temporal reasoning (CATER; [5]), and causal visual reasoning (ACRE;[14]). Our approach outperforms prior state-of-the-art particularly designed to perform one of these tasks by a large margin.

2. Look, Remember, Reason

To allow visual reasoning by exploiting the highly expressive large language models, we propose a novel *"Look, Remember, Reason"* framework. To address the challenges presented by visual reasoning problems, we propose rationales obtained from multimodal signals. Unlike prior work [15, 18], our rationales additionally include low-level visual surrogate tasks expressed in natural language crucial for visual reasoning tasks.

2.1. Auto-regressive Pipeline

Inspired by the success of auto-regressive models in reasoning tasks [2], we formalize our LRR model in the auto-regressive framework. Our LRR model (Fig. 2) with parameters θ receives an interleaved stream of visual input, $\mathbf{I} = (\mathbf{v}_1, \ldots, \mathbf{v}_{t_v})$, *e.g.*, a sequence of images of length t_v, along with (tokenized) text $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{t_s})$ of length t_s. The tokenized text includes the rationales and answers to visual reasoning problems. We train the model by maximizing log-likelihood of the next token,

$$\log(p_\theta(\mathbf{S})) = \sum_{t'_s} \log(\mathbf{s}_{t'_s} | \mathbf{s}_1, \ldots, \mathbf{s}_{t'_s - 1}, \mathbf{v}_1, \ldots, \mathbf{v}_{t'_v}) \quad (1)$$

where, $(\mathbf{v}_1, \ldots, \mathbf{v}_{t'_v})$ is the interleaved visual input sequence upto the text token $\mathbf{s}_{t'_s}$. The backbone of our model consists of an off-the-shelf LLM. We use models from the OPT family [17], but verified that similar performance can be achieved using other pre-trained models [12]. The parameters are initialized from pre-trained LLMs, which allows us to exploit their existing reasoning capabilities. While the LLMs we use as backbone are trained on text only, visual reasoning relies on the extraction of visual information about spatial and temporal relationships between objects in the scene. Therefore, in our multi-modal setup, visual information $\mathbf{I}$ needs to be mapped to the text-based representation space of the LLM. The key challenge here is that in comparison to text tokens, images are highly information dense – reflected in the popular adage "An image is worth a thousand words". To this end, we insert cross attention layers in between the transformer layers in the LLM (Fig. 2) and exploit grid-based features from off-the-shelf CNNs as they are better at preserving low-level visual information compared to encoders such as Perceiver [6] or CLIP [11] used by state-of-the-art multi-modal LLMs [1, 9]. This is because compared to Perceiver [6] or CLIP [11], grid-based features from CNNs such as ResNet do not include global pooling operations which destroy low-level visual information. (see supplementary for details)

2.2. Rationales with Surrogate Tasks

In our LRR model, we leverage the flexibility of LLMs to express diverse low-level visual tasks through language in

	Datasets						
	CLEVR	CATER: Static Camera		CATER: Moving Camera		ACRE: Comp	ACRE: Sys
Method	Acc↑	Top 1↑	Top 5↑	Top 1↑	Top 5↑	Acc↑	Acc↑
SOTA	**99.7**	79.7	95.5	59.7	90.1	91.7	93.9
LRR (w/o Rationale)	51.4	61.7	82.4	49.3	65.8	89.7	90.2
LRR (Fine-tuned)	97.9	**85.1**	96.2	75.1	91.9	**99.3**	**99.0**
LRR (Joint)	97.3	83.7	**96.5**	**75.2**	**92.8**	98.9	98.7

Table 1. Evaluation of our LRR model.

a generalized setup. Consider the visual reasoning problem from ACRE in Fig. 1, which requires low-level skills of object recognition and re-identification. We introduce the surrogate task of recognizing each object class and assigning each object a unique identifier across context trials in the rationale. Including low-level visual tasks in the rationale has the additional benefit that the solutions to these tasks remain within the context window of the LLM so that they are in fact "remembered" by the LLM and can be exploited to "reason" and solve subsequent tasks. The experiments section provides practical details on rationale construction.

3. Experiments

We now evaluate our model on a diverse range of visual reasoning tasks, including: 1. ACRE [14] which focuses on the problem of causal discovery. 2. CLEVR [7] which focuses on spatial reasoning, and 3. CATER [5] which focus on temporal reasoning.

Models and training details. We focus on the OPT family of LLMs [17], particularly OPT-1.3B. We train our LRR model on a single Nvidia A100 GPU. We used a ResNet-101 as the vision backbone across all tasks (see supplementary).

3.1. Rationale Construction

We next describe how rationales are constructed for training our LRR model (see supplementary for more details).

ACRE. The ACRE dataset [14] focuses on evaluating the performance of vision systems on the problem of causal induction. Specifically, the dataset focuses on the problem of causal discovery using "blicket" detection experiments, originally administered to children. The experiment involves a series of context trials, in which various (combinations of) objects are placed on the blicket detector, and subjects are shown whether the detector is activated. They are then asked which objects or (novel) combinations of objects would activate the machine. The key low-level visual challenge in the ACRE dataset is to identify objects in the context trials and to detect whether the blicket machine is activated. Therefore, we design the rationale with the surrogate tasks of object recognition and re-identification across the context trials. The rationale for each context trial describes the objects present and also assigns an unique integer ID to allow for re-identification. Additionally, the rationale also identifies state of the blicket machine (on/off) (see Fig. 1).

CLEVR. The training set of CLVER is annotated with functional programs with sub-routines that decompose questions into simpler low-level object recognition and spatial reasoning tasks such as object counting and searching for objects based on spatial positions or materials, among others – operations necessary to solve the visual reasoning problem. We convert these sub-routines into rationales with surrogate tasks corresponding to these low-level object recognition and spatial reasoning tasks.

CATER. The CATER (Compositional Actions and TEmporal Reasoning) dataset is designed to test the ability to recognize compositions of object movements that require long-term temporal reasoning. Similar to [3], we focus on the hardest task from the CATER dataset, *i.e.*, adversarial target tracking under occlusion and containment. This task is posed as a classification problem over a 6×6 grid. We decompose the final grid classification problem into a sequence of simpler problems, using rationales with multi-target tracking as a surrogate low-level visual task. The rationale contains the grid positions of the snitch at every video frame. Following the paradigm of "Look, Remember, Reason" we include the surrogate task of tracking the medium and large cones in the scene, as these objects can occlude the snitch. With our rationale, the predicted intermediate grid positions of the objects of interest, *e.g.*, the snitch and cones, are "remembered" by the LLM and can be used to reason about the final position of the snitch in case of recursive containment by the cones.

3.2. Quantitative Evaluation

We provide an overview of our results in Table 1, with dataset specific baselines and ablations in the supplementary. We compare to the state of the art (SOTA) models on CLEVR: [8, 13], CATER: [3, 16] and ACRE: [3]. We evaluate two versions of our LRR model: LRR (Fine-tuned) and LRR (Joint). They are fine-tuned per dataset and jointly trained across CLEVR, CATER and ACRE datasets respectively. Additionally, we also consider a version of our LRR

model trained without rationales (w/o Rationale). Our LRR (Fine-tuned) model outperforms the state-of-the-art by 7.6% and 5.1% on the compositional and systematic splits of the ACRE dataset and; by 5.4% Top-1 accuracy on static camera and 15.4% Top-1 accuracy on moving camera splits of the CATER dataset. Further, the performance of our LRR model is comparable to the state-of-the-art with task-specific architectures on CLEVR. Our jointly trained LRR model for the first time shows performance comparable to the dataset specific fine-tuned variants on such diverse visual reasoning tasks. This shows the ability of our LRR model to adapt to diverse visual reasoning tasks encountered in the real-world.

4. Conclusion

We show that off-the-shelf LLMs can solve complex visual reasoning tasks when supervised with rationales with surrogate visual tasks and equipped with top-down visual attention. We exploit the flexibility of LLMs in language modeling, which allows us to express diverse low-level visual tasks, *e.g.*, recognition, tracking, and re-identification, in the form of language. The use of off-the-shelf LLM and vision backbones allows our model to be readily applicable across diverse tasks.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

[3] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt M. Botvinick. Attention over learned object embeddings enables complex visual reasoning. In *NeurIPS*, 2021.

[4] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. *CoRR*, abs/2303.03378, 2023.

[5] Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for compositional actions & temporal reasoning. In *ICLR*, 2020.

[6] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021.

[7] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[8] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.

[9] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *CoRR*, abs/2301.13823, 2023.

[10] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[12] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022.

[13] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *NeurIPS*, 2018.

[14] Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. ACRE: abstract causal reasoning beyond covariation. In *CVPR*, 2021.

[15] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *CoRR*, abs/2303.16199, 2023.

[16] Shiwen Zhang. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning. *CoRR*, 2022.

[17] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022.

[18] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923, 2023.