Report for ImageCoDe Challenge: Leveraging Chain-Of-Thought Prompting Techniques

Wu Ding
Student, McGill University
Department of Computer Science
wu.ding@mail.mcgill.ca

Harrison Finkelstein-Hynes
Student, McGill University
Department of Computer Science
harrison.finkelstein-hynes@mail.mcgill.ca

Abstract

Identifying a scene based on a description of slight differences from its neighbors is an important task for both natural language understanding and image-processing. In humans, this ability to discern an image based on the contextual clues is one that is innate and natural. However, as shown in the original ImageCoDe paper, this task is far from trivial for even the state-of-the-art machine learning models. The challenge described in that paper is one of correctly identifying a target image from a set of 10 images with minimal distinctions using a short description.

We will implement prompting methods on Large Language Models with vision based aspects using different chain-of-thought reasoning techniques. These techniques include breaking down the task into simpler atomic subtasks (DDCoT) and listing the similarities and differences between images (CoCoT) before having to identify the target. We hope that attacking the task in this fashion will result in better performances of our models in the ImageCoDe challenge. If we do not succeed, we hope to explain our shortcomings in order to further advance our general comprehension of the problem.

1 Introduction

NLP research has seen a growth of solutions to problems that involve no training (Liu et al., 2023). While pre-training and fine tuning models can be successful, simply prompting language models during inference time has been shown to be sufficient for many tasks (Zheng et al., 2023b). Finding successful prompts has become an area of research itself with big breakthroughs (Kojima et al., 2022; White et al., 2023; Ekin, 2023). Few-Shot Prompting, provides question and answer pairs to a model to point a model in the right direction (Brown et al., 2020). Chain-Of-Thought (CoT) prompting takes this a step further, the question/answer pairs now include a section that explains the reasoning behind the answer (Wei et al., 2022). CoT prompting has been successfully applied to vision-language classification tasks, achieving gains over CLIP (Radford et al., 2021) based models, and achieving higher scores specifically in areas relating to pragmatics (Wu et al., 2023). Contrastive CoT (CoCoT) adds a layer again for vision-language classification tasks, by prompting the model to articulate not just its reasoning, but the similarities and differences in the images (Zhang et al., 2024). In two-turn CoCoT, the model is asked questions about the images based on the similarities and differences it noticed.

Model	Accuracy
Human Performance	90.8
NDCR-v2	34.1
ALBEF-finetuned	33.6
NDCR	32.6
Baseline: ContextualCLIP	29.9
DCIG	28.4

Table 1: Current ImageCoDe Leaderboard from https://mcgill-nlp.github.io/imagecode/

The ImageCoDe benchmark focuses on understanding the nuances of language and its relation to images (Krojer et al., 2022). It tests models on their ability to comprehend complex descriptions of images that require an understanding pragmatics (understanding implied meaning), temporality (the sequence of events), complex descriptions, and subtle visual details. This challenging task goes beyond image recognition, instead models must build a pick out the correct image from a minimally contrastive set, on the basis of a dense linguistic direction. ImageCoDe highlights the limitations of current language and vision models in grasping these complex connections. Previous attempts at the benchmark have failed to come close to human performance.

In this research we will explore the effectiveness of these prompting techniques to solve the ImageCode task. We believe that applying these tried and true prompting techniques to this problem will result in gains over the CLIP-contextual models. We believe that the complexity of the ImageCode task will benefit from the generalized context of state of the art large multi-modal models (LMM). We hope to show that improvements on the ImageCode task by leveraging the effectiveness of prompting large multi-modal models.

2 Related Works

2.1 ImageCoDe

In their original paper (Krojer et al., 2022), the authors described their dataset and set benchmarks for the new ImageCoDe task. State-of-the-art language-and-vision models such as ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020) and CLIP (Radford et al., 2021) were trained and evaluated on the ImageCoDe task. Different methods, such as providing the same contextual batch or combining temporal embeddings during training, provided different results. In the end, the benchmark was set using the CLIP encoder trained with contextual modules and temporal embeddings. As shown in Table 1, the accuracy using this technique was 29.9%. While there have been attempts at improving the state-of-the-art accuracy (Ou et al., 2023; Li et al., 2023), we are still far from matching the performance of humans.

2.2 Chain-of-thought

Research has shown that prompting large language models with the right text prompt evokes different behaviors that resemble reasoning (Wei et al., 2022; Kojima et al., 2022). Prompts can range from providing a few-shot chain-of-thought (CoT) reasoning (Wei et al., 2022) to simply appending “Let’s think step by step” (Kojima et al., 2022) to a zero-shot prompt. This type of simple addition, when passed the LLMs, can improve the performance on certain tasks involving reasoning on text. Likewise, research has been conducted to implement this CoT reasoning on tasks combining text and images. Breaking the problem up into several sub-problems drastically improved the performance of LLMs such as GPT-4V evaluated on the Winoground benchmark (Wu et al., 2023), which is a task consisting of selecting the correct image from a text description, or vice-versa (Thrush et al., 2022). Similarly, decomposition of the task into logical steps, either in the form of a python program (Surís et al., 2023) or pseudocode (Gupta and Kembhavi, 2022) can improve on visual tasks including visual grounding and compositional visual QA. These improvements were done without any training on those specific tasks (Gupta and Kembhavi, 2022), which is very exciting as we hope to do the same on the ImageCoDe challenge.

2.3 Contrasting Chain-of-thought

A special type of chain-of-thought reasoning is called contrastive chain-of-thought reasoning (CoCoT), which involves first probing the LLM for similarities and differences between the set of inputs, which in our case would be images, and then completing the task using the deducted information (Zhang et al., 2024). This type of CoT, as explained by (Zhang et al., 2024), differs from its counterparts Duty-Distinct Chain-of-Thought (DDCoT) (Zheng et al., 2023a) and Compositional Chain-of-Thought (CCoT) (Mitra et al., 2023) through the specific nature of the prompt in question and exactly what part of the task is being broken down. CoCoT’s prompt requests specific information from the set of images that will help us in our ImageCoDe task, especially when prompting the differences, as they are precisely what the ImageCoDe descriptions are based on (Krojer et al., 2022). Other models have been trained regarding the description of differences between similar images (Jhamtani and Berg-Kirkpatrick, 2018), but the advantage of using the CoCoT method is the flexibility it provides from the lack of required training. It was shown that CoCoT provided state-of-the-art performance (Zhang et al., 2024) across different datasets while using several different large multimodal models (LMMs) including OpenFlamingo (Awadalla et al., 2023), MMICL (Zhao et al., 2023), GEMINI (et al., 2023) and GPT-4V (Yang et al., 2023). Thus, CoCoT allows us to test the work of more recent models.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. arXiv:2005.14165. arXiv. ArXiv:2005.14165 [cs].

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Sabit Ekin. 2023. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices.

Gemini Team et al. 2023. Gemini: A family of highly capable multimodal models.

Tanmay Gupta and Aniruddha Kembhavi. 2022. Visual programming: Compositional visual reasoning without training.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, page 22199–22213. Curran Associates, Inc.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yunxin Li, Baotian Hu, Yuxin Ding, Lin Ma, and Min Zhang. 2023. A neural divide-and-conquer reasoning framework for image retrieval from linguistically complex text.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a CLIP listener for contrastive captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1904–1917, Toronto, Canada. Association for Computational Linguistics.

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. 2021. Learning transferable visual models from natural language supervision.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C. Gee, and Yixin Nie. 2023. The role of chain-of-thought in complex vision-language reasoning task.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. (arXiv:2401.02582). ArXiv:2401.02582 [cs].

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023a. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models.

Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, and Ram Nevatia. 2023b. Large language models are good prompt learners for low-shot image classification.

A Appendix

We will be working with the dataset collected for the ImageCoDe challenge (Krojer et al., 2022). This dataset consists of 94020 images and 21202 image descriptions, all of which were annotated by humans. Each description describes, with the least amount of information, a single correct image from a set of 9 other similar, but incorrect, ones from the grouped set. The images themselves can be classified into two categories: static pictures, which are images obtained from a dataset of independent images, and video frames, which are images obtained by taking frames of different videos. Thus, the images obtained from videos can appear more similar to one another and can also incorporate a temporal aspect.

Here is an example of an instance of data:

```
{"image_set": "video-storytelling
   -videowedding_de8dLXvgV-I-
   shot6_0",
"image_index": "8",
"description": "The flowers the
   woman in the teal strapless
   dress is carrying are
   completely obscured by the man
     in the black shirt's head. "}
```

The image_set refers to the set of 10 similar images, the image_index refers to the index of the target image within that set, and the description refers to the differences of that image in the goal of identifying it.

Finally, our evaluation metric will be the accuracy of our models' identification of the correct target image. We will compare our results with the ones on the ImageCoDe leaderboard, notably with the baseline set by ContextualCLIP from the original ImageCoDe paper (see Table 1).