

Multimodal chain-of-thought

Anonymous Authors¹

Abstract

Language models (LMs), particularly large-scale models (LLMs), have demonstrated remarkable capabilities across various tasks due to a reasoning process, called chain-of-thought (CoT), emerged in them. However, CoT, while improving predictive quality, comes at the cost of increased inference time. In this project, we propose a novel approach to mitigate this trade-off by introducing compressed CoT representations. Inspired by successful techniques in light-weight multimodal LLM extensions, we employ a two-step process: first, training compressed data representations using effective compressing methods i.e quantized autoencoders, and then fine-tuning pre-trained LLMs with an extended vocabulary incorporating "tokens" from different modalities. We present our methodology and discuss the potential impact on improving both efficiency and effectiveness of LLMs. Experimental results are produced on two different multi-modal datasets: (1) ScienceQA multi-modal dataset with text and images; (2) Sk3D dataset with images, depth maps and meshes. The second dataset is additionally annotated with detailed instructions suitable for training LLMs with CoT. Please find out project on Git: <https://github.com/BohrG/MMDCOT>

1. Introduction

Large-scale language models (LLMs) have revolutionized natural language processing (NLP) tasks with their ability to generate coherent and contextually relevant text. In the realm of Language Modeling, strategies such as Chain-of-Thoughts have demonstrated the ability to enhance the predictive capabilities of LLMs across various tasks. However, the use of such methods comes at the cost of increased

inference complexity.

The recent works on Multimodal LLMs and Chain-of-Thought inspired us to investigate the capabilities of injecting qualified reasoning in LLMs via training on diverse data modalities. So, Multimodal CoT (Zhang et al., 2024) work expands on the Chain-of-Thought approach by integrating multimodal data, enabling the model to process and reason across textual, visual, and auditory inputs what allows for a more nuanced understanding and generation of responses that reflect a deeper comprehension of mixed-media content. AnyGPT framework (Zhan et al., 2024) proposes a novel approach to building generative models that are not limited to text but can seamlessly incorporate and generate across different modalities, including images and speech. Authors demonstrate that such an approach not only broadens the applicative potential of LLMs, but also improves the model's robustness and reasoning by leveraging cross-modal synergy during training.

Several scholarly works have highlighted the observation that language models possess redundant capacity which remains unused. In response, (Ashkboos et al., 2024) advocate for a methodological approach to alleviate this redundancy by employing orthogonal transformations, subsequently followed by the removal of dormant rows and columns from the weight matrices. A series of empirical investigations demonstrated that the embedding dimensionality of the resultant model is reduced by approximately 10-30%. Furthermore, the evaluation metrics, particularly perplexity, exhibit minimal deterioration, notably in the case of large-scale language models such as LAMA-70B and OPT-66B.

Drawing inspiration from these concepts in our project, we propose the idea to develop compressed representations of chains of thought facilitating the substitution of the LLM reasoning process with a reduced set of distinctive tokens. We propose a two-stage framework by training compressed data representations using techniques such as quantized autoencoders to reduce inferential complexities, then fine-tune pre-trained LLMs with an expanded vocabulary to incorporate multimodal "tokens". We introduce Sk3D dataset with additional annotations for 3D reconstruction from multiple modalities. Thus, we prepare a dataset of mixed-modality data and fine-tune LLM Incorporate point clouds as an additional modality, leveraging their spatial data characteristics

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

to enrich the model’s understanding and predictive capabilities.

Expected advantages of our method: (1) improved efficiency due to compressed chain-of-thought reasoning; (2) broadened modalities landscape by introducing 3D data input along with 2D visual and text data on novel dataset.

2. Related Work

Recent advancements in language model (LM) prompting techniques have unveiled the potential for improved performance by encouraging models to engage in a form of deliberative thinking. While conventional zero-shot prompts aim for immediate responses, recent articles suggest that prompting LMs to “think aloud” about the problem space can lead to more accurate outcomes. (Wei et al., 2022) (Kojima et al., 2022) This concept, often referred to as “chain-of-thoughts” (CoT) entails guiding the LM through a step-by-step reasoning process similar to human deliberation. The addition of a simple “Let’s think step-by-step” prompt enabled the PaLM language model to outperform humans on 10 out of 23 tasks in the Big-Bench benchmark. (Suzgun et al., 2022)

An alternative method, termed Quiet-STaR, enables a model to engage in silent contemplation at each token, with a trained distribution optimized for utility. The authors proposed an idea to generalize CoT prompting, approximating it to the semblance of human thought process (not expressing everything one thinks). To achieve this, the authors suggested allowing LLM to sometimes “think” — the start and end of such thought are defined by trainable reinforcement learning (RL) special tokens, while the thought itself often consists of gibberish, reducing the perplexity of the subsequent text. Unlike CoT, there is no requirement to write coherently here, so what is generated inside the thought is not always interpretable. Nevertheless, this significantly boosts metrics and practically does not slow down inference, as generation occurs in parallel.

Data tokenizers play a key role in processing and understanding different data modalities in LLMs. For common modalities such as text, images, and speech, well-established tokenizers have been developed, enabling the efficient training of models on diverse datasets i.e AnyGPT. (Zhan et al., 2024) However, tokenizing more complex data types that are crucial for deep understanding of a 3D environment i.e 3D point clouds or meshes, presents unique challenges due to their intricate structures and the spatial relationships. Recent innovations in this area include PointBERT (Yu et al., 2022) and MeshGPT (Siddiqui et al., 2023), which have presented innovative approaches for tokenization of 3D data forms. PointBERT proposes a tokenizer for point clouds – a 3D data representation in a form of a discrete set of data points in space. This tokenizer is based on a discrete Varia-

tional AutoEncoder which generates discrete point tokens containing meaningful local information and is pre-trained on ShapeNet dataset. MeshGPT is aimed to process meshes – a different 3D representation, which consists of vertices and faces. MeshGPT learns a vocabulary of latent quantized embeddings, using graph convolutions. This approach not only preserves the geometric and topological properties of the meshes, but also facilitates the generation and manipulation of 3D shapes using language model-like techniques. In this project, apart from working with well-known modalities, we aim to dive deeper into the 3D domain and utilize tokenizers proposed for 3D data modalities.

3. Methods

3.1. Problem Formulation

The primary objective of this research is to devise a compressed representation methodology for the reasoning process within LLMs. This approach aims to facilitate the substitution of the conventional CoT reasoning generation mechanism with a reduced set of specialized tokens. Compressing the intricate reasoning process into a more concise form, the proposed methodology seeks to enhance the efficiency and scalability of LLM-based reasoning tasks while maintaining or even augmenting their predictive capabilities.

The overall procedure is composed of three main components:

- **Train Compressed Data Representation.** Initially, a compressed data representation is trained. One possible way that we propose in this project is to use a quantized autoencoder to discretize the latent space, leading to even more compact representations.
- **Fine-tune pre-trained LLM.** Subsequently, the LLM is fine-tuned using the compressed data representation. This fine-tuning process involves extending the vocabulary of the LLM to include “tokens” representing different modalities. These modalities could include various types of data such as text, images, audio, as well as 3d point clouds. By incorporating tokens representing different modalities, the LLM becomes capable of processing and generating outputs based on a diverse range of inputs, thereby enhancing its versatility and applicability.

4. Experiments

4.1. Dataset

The aim to build a deep multi-modal reasoning in LLMs requires enough diverse data with text and other data modalities. We assess the effectiveness of our approach using the

ScienceQA benchmark (Lu et al., 2022), renowned as the pioneering large-scale multimodal science question dataset, which meticulously annotates answers with extensive lectures and explanations. Comprising 21,000 multimodal multiple-choice questions, the dataset exhibits significant domain breadth, spanning three subjects, 26 topics, 127 categories, and 379 skills.

Another dataset, **Sk3D** (Voynov et al., 2023) is one of the recently proposed benchmarks for 3D reconstruction from multiple modalities. It provides more than 1.4M data samples of different modalities, such as images, depth maps from different types of scanners and meshes for 110 real-life objects. Yet, this dataset lacks text instructions that can be used as annotations for LLM finetuning. So, this dataset will be additionally annotated with an automated approach using GPT-like models to generate detailed instructions.

4.2. Citations and References

Software and Data

Acknowledgements

References

- Ashkboos, S., Croci, M. L., Nascimento, M. G. d., Hoefler, T., and Hensman, J. Slicept: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.
- Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., and Nießner, M. Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv preprint arXiv:2311.15475*, 2023.
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Voynov, O., Bobrovskikh, G., Karpyshev, P., Galochkin, S., Ardelean, A.-T., Bozhenko, A., Karmanova, E., Kopanov, P., Labutin-Rymsho, Y., Rakhimov, R., Safin, A., Serpiva, V., Artemov, A., Burnaev, E., Tsetserukou, D., and Zorin, D. Multi-sensor large-scale dataset for multi-view 3d reconstruction. *CVPR*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. *CVPR*, 2022.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Zhang, Z., Zhang, A., Li, M., hai zhao, Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *ICLR*, 2024.