Efficient Inference for Large Multimodal Models

Chaoqun Yang
Tsinghua University

Yuanda Zhang
Tsinghua University

Xiying Huang
Tsinghua University

Abstract

Large Multimodal Models (LMMs) have achieved notable success in visual instruction tuning, yet their inference is time-consuming due to the auto-regressive decoding of Large Language Model (LLM) backbone. Speculative Decoding (SD) has proven effective for lossless auto-regressive decoding acceleration by a *draft-then-verify* mode. In this work, we explore the application of speculative decoding to boost the inference efficiency of LMMs, with a particular focus on leveraging useful information generated during the LMM's processing in the context of multimodal tasks, such as vision embeddings, hidden states, and key-value (KV) caches. Concurrently, we try to develop alignment techniques between the target model and the draft model in this scenario, with the objective of improving the acceleration effect as much as possible. We anticipate achieving a speedup ratio of over 2x. Code and model will be released in the near future.

1 Background

The rapid development of large models is exerting a profound impact on the whole world. These models have demonstrated remarkable capabilities in various domains, including text generation, visual understanding, question answering, logical reasoning and video generation [1, 3, 14, 15]. These large models are expected to move towards artificial general intelligence. However, as the scale and complexity of large models increase, so does the computational cost associated with their inference [6, 16]. This issue is particularly pronounced in large multimodal models, which integrate visual data with textual information, further exacerbating the latency and resource demands [20].

The problem of costly inference in LMMs is not merely a theoretical challenge; it is deeply rooted in practical application scenarios that require real-time responses. Addressing this issue has far-reaching implications, as it not only reduces latency, thereby improving user interactions with AI systems, but also conserves computational resources, thus decreasing energy consumption associated with running these models.

Recently, speculative decoding emerges as a prospective method for accelerating inference without loss of accuracy, which has demonstrated effective utility in large language models [7, 17, 19]. However, research on its application in accelerating LMMs is still relatively scarce. Several research questions in this domain await resolution. How much gain can Speculative Decoding provide for the inference acceleration of LMMs? How can the alignment between the draft model and the target model be achieved in a multimodal context? What information can be leveraged and how should it be utilized during this alignment process? What are the practical deployment costs of this acceleration framework, and is it acceptable in real-world applications?

Our research aims to explore these questions, offering both theoretical and practical contributions to the acceleration of LMMs inference.

Preprint. Under review.

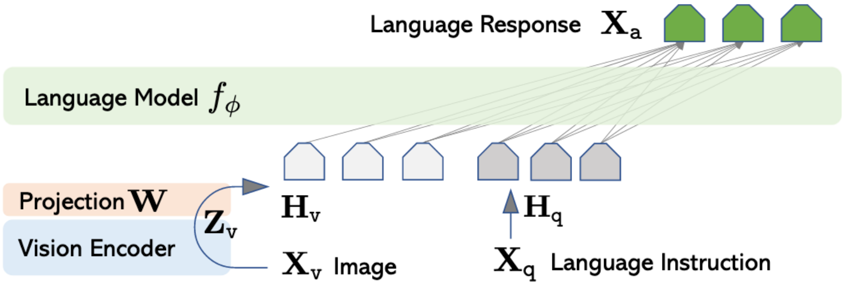


Figure 1: Instruction-following Large Multimodal Model network architecture [12].

2 Define

In the context of this paper, we define speculative decoding as a method to accelerate the inference process of LMMs by parallelizing the verification of speculatively generated tokens. Formally, given an input with visual encoding X_V and text tokens X_T, where $X_V = (x_1^V, x_2^V, ..., x_m^V)$, $X_T = (x_1, x_2, ..., x_n)$, the speculative decoding process can be described as follows:

Drafting Stage: A draft model M_d generates a sequence of speculative tokens $S' = (x'_1, x'_2, ..., x'_\gamma)$ based on the input embedding (X_V, X_T).

Verification Stage: The target model M_T verifies the speculative tokens S in parallel, getting the real token sequence $S = (x_1, x_2, ..., x_k)$.

3 Related Work

3.1 Large multimodal models

The typical Large Multimodal Model considered in this context is one that is adept at following instructions, often referred to as an Instruction-following Large Multimodal Model [11, 12], such as LLaVA [12], Qwen-VL [2], InstructBLIP [4], GPT-4 [1]. Its common architecture consist of 1) a vision encoder to encode the input image, 2) a vision-language connector to convert the image encodings to language model embeddings, and 3) a language model backbone (Figure 1).

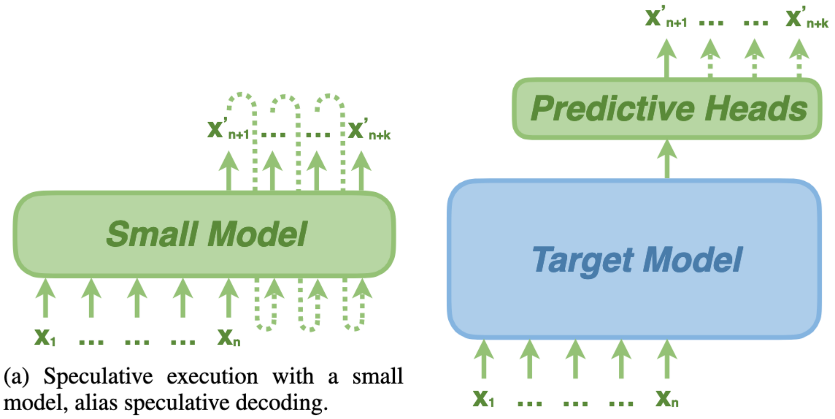
3.2 Speculative decoding

Speculative Decoding uses the idea of draft-then-verify to fully leverage the parallel processing capabilities of GPU [8]. Specifically, this method initially utilizes a draft model, typically a small and rapidly executing model, to generate multiple draft tokens. Subsequently, the target model verifies these draft tokens in a single, parallel operation, reducing the number of auto-regressive decoding steps of the target model, thereby accelerating the overall inference process [17]. Recent advancements that design predictive heads to leverage target LLM's contextual information, such as hidden states, have shown significant practical improvements [9, 18, 21]. Unlike independent small models, predictive heads are directly integrated onto the target model, allowing them to leverage the information and context inherent in the larger model (Figure 2).

Regarding LMMs and SD, there is no dedicated research on speculative decoding algorithm specifically for LMMs, with only preliminary and rudimentary experiments conducted thus far [5], which has inspired our research.

4 Proposed Method

Motivated by the success of speculative decoding in accelerating LLM inference, and observing the auto-regressive decoding pattern in LMM for its language model backbone, we propose to apply this approach to LMMs to enhance their inference efficiency. The following is our planned experimental settings.



(a) Speculative execution with a small model, alias speculative decoding.

(b) Speculative execution with multiple predictive heads, alias blockwise parallel decoding.

Figure 2: Predictive heads v.s. small models [19].

4.1 Experimental settings

Models and datasets. Refer to the previous work on speculative decoding, we intend to conduct experiments on LLaVA-7B and LLaVA-13B [12], encompassing the common sizes of current mainstream LMMs.

Datasets and tasks. We plan to evaluate our method across multiple tasks including mixed generic tasks (conversation, detailed description, and complex reasoning) on LLaVA-Bench In-the-Wild [12] dataset, Image captioning task on images from COCO [10] dataset, and chain-ofthought (CoT) reasoning on Science QA [13] dataset.

Metrics. Like other work about speculative decoding, we assess acceleration effects using the following metrics:

- **Walltime speedup ratio:** The actual test speedup ratio relative to vanilla auto-regressive decoding.
- **Average acceptance length:** The average number of accepted tokens per block (or target model forward) for a block size γ (or speculative steps).
- **Ave acceptance rate:** The average of the ratio between the number of accepted tokens and the number of speculative tokens.
- **Token rate:** The average number of generated tokens per second.

4.2 Baseline

Refer to the previous work [5], we intend to establish the following baseline:

- 115M pretrained-LLaMA as draft model.
- 115M finetuned-LLaMA as draft model.
- finetuned-LLaVA as draft model.
- the language model part from finetuned-LLaVA as draft model.

4.3 Our Method

Predictive heads design. We propose to design a lightweight network architecture, which is tailored to effectively capture pertinent information from image embeddings and hidden states, predicting the tokens to be generated.

Alignment method design. We propose to design a specialized knowledge distillation technique, which aims to achieve effective alignment between the predictive heads and the target model within the context of speculative decoding for LMMs.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[4] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, pages 49250–49267. Curran Associates, Inc., 2023.

[5] Mukul Gagrani, Raghavv Goel, Wonseok Jeon, Junyoung Park, Mingu Lee, and Christopher Lott. On speculative decoding for multimodal large language models. *arXiv preprint arXiv:2404.08856*, 2024.

[6] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[7] Mahsa Khoshnoodi, Vinija Jain, Mingye Gao, Malavika Srikanth, and Aman Chadha. A comprehensive survey of accelerated generation techniques in large language models. *arXiv preprint arXiv:2405.13019*, 2024.

[8] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

[9] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[13] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[14] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[16] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.

[17] Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. *arXiv preprint arXiv:2401.07851*, 2024.

[18] Bin Xiao, Chunan Shi, Xiaonan Nie, Fan Yang, Xiangwei Deng, Lei Su, Weipeng Chen, and Bin Cui. Clover: Regressive lightweight speculative decoding with sequential knowledge. *arXiv preprint arXiv:2405.00263*, 2024.

[19] Chen Zhang, Zhuorui Liu, and Dawei Song. Beyond the speculative game: A survey of speculative execution in large language models. *arXiv preprint arXiv:2404.14897*, 2024.

[20] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

[21] Lefan Zhang, Xiaodan Wang, Yanhua Huang, and Ruiwen Xu. Learning harmonized representations for speculative sampling. *arXiv preprint arXiv:2408.15766*, 2024.