Qeios

Peer Review

Review of: "Draft Model Knows When to Stop: A Self-Verification Length Policy for Speculative Decoding"

Junhui Li[1]

1. Independent researcher

This paper addresses speculative decoding (SD) and introduces a new method, SVIP, which is a difficulty-aware dynamic draft length policy for draft generation in SD. SVIP determines whether to continue generating the next draft token or stop, based on the entropy of the predicted token. A threshold is used to make this decision. The paper provides a solid motivation for the method, along with a theoretical analysis. Experimental results across various LLMs and benchmarks demonstrate the effectiveness of the proposed SVIP.

Strengths:

1. SVIP is well-motivated, simple, and effective. The paper also presents a thorough theoretical analysis to support it.

2. SVIP is a plug-and-play dynamic draft length policy, which offers flexibility as it is training-free and can be easily adapted to an auto-regressive draft model in SD systems.

3. The paper provides comprehensive experimental results and insightful discussions that enhance our understanding of how SVIP achieves significant speedup performance.

Weaknesses:

While the paper is strong overall, some clarifications would improve its presentation:

1. In Figure 2, it is unclear which SP system is being studied (is it the baseline draft length policy or SVIP?). Also, what is "root" draft entropy in Figure 2 (top)?

2. Figures 6 and 7 show that, despite having lower acceptance rates, the Constant and Heuristic baselines sometimes accept more tokens than SVIP, resulting in less time for the target models. A

 doi.org/10.32388/RQVY8A

discussion on the individual time consumption of the target model versus the draft model would provide clearer insight into how SVIP outperforms the two baselines.

3. In footnote 1, the term q(x_n) can also be included in the explanation.

Typos:

In Section 2.1, should "x_{n+j}" be "x_{t+j}"?

Declarations

Potential competing interests: No potential competing interests to declare.